“빅데이터 아키텍처 수립 평가 과제” 로 알아보는

하둡 기반 빅데이터 시스템

Encore DB반

나세화

과제 내용

Condition A

- 3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것

  A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  C. Data Block Size는 32MB로 설정
  D. Block Replication(복제)는 2로 설정
  E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공

  A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  B. Hive JDBC를 제공할 것
  C. 데이터가 존재하는 테이블을 3개 이상 만들 것

- Spark 서비스 제공
- Zeppelin notebook 서비스 제공

과제 내용

Condition B

- 각 Host는 아래의 사양을 만족해야 한다.
  - A. 각 Host의 Disk Size는 48 ~ 64GB로 설정한다.
  - B. 각 Host의 CPU Core는 2~4개로 설정한다.
  - C. 각 Host의 RAM Size는 4.8GB 이상으로 설정한다.

- Root는 패스워드가 없어야 한다.
  - A. HDP 설치 시 ssh private key 입력을 통한 인증(키교환)으로 Host를 등록해야 한다.

- 클러스터는 외부에서 접속 가능하도록 가상 브릿지 네트워크를 이용한 통신 구성을 해야한다.
  - A. 네트워크 대역은 상관 없음

- 방화벽 서비스는 disable
- 서비스 연결 시 host명 return 문제를 해결해야 한다.

과제 내용

Condition C

- HDP에 포함된 service의 용도 및 역할을 설명한다.
- 시스템 구성도를 작성한다. (VM(Virtual Machine), 네트워크, service 구성 등 명시)

- 상기 조건들을 모두 확인할 수 있는 가이드를 제공해야 한다.

  A. HDP 관리 페이지 접속 방법
  B. Hive 접속 방법
  C. Spark History 서버 접속 방법
  D. Zeppelin notebook 접속 방법

과제 내용

Option

- OS 종류는 상관 없음
- VM, Docker, Native 모두 상관 없음
- 위 조건을 만족 시키기 위해 별도의 콤포넌트, 서버 등을 추가 하여도 상관 없음.
- 하둡 에코시스템 각 구성요소들을 별개로 설치 하여도 무방함.
- 모든 설정은 '설정 값 ' 을 기준으로 평가.
  즉, 설정이 적용되어 설정대로 동작 하는지를 평가 하는 것이 아니라
  '정확한 설정 파일에 정확한 설정 값' 이 들어가 있는지만 확인.
- 상기 조건만 만족 할 경우 부수적인 것들은 상관 없음.
- 최종 산출물에 error, warning 등이 있더라도 상기 조건에 해당하지 않으면
  평가에 반영치 않음.

Condition A, Condition B만 알아보자!

Condition A

Hadoop?

- 하둡이란 무엇일까? -

"대용량 데이터를 **분산처리** 할 수 있는
자바 기반 오픈소스 프레임워크"

Hadoop?

- 하둡이란 무엇일까? -

HDFS

MapReduce

분산 파일 **저장** 시스템

분산 파일 **처리** 시스템

"대용량의 데이터를 **HDFS**에 저장하고,
MapReduce로 처리하여 분산처리 한다!"

Hadoop?

- 하둡이란 무엇일까? -

HDFS

MapReduce

분산 파일 **저장** 시스템

분산 파일 **처리** 시스템

HDFS?

"대용량의 데이터를 **HDFS**에 저장하고,
MapReduce로 처리하여 분산처리 한다!"

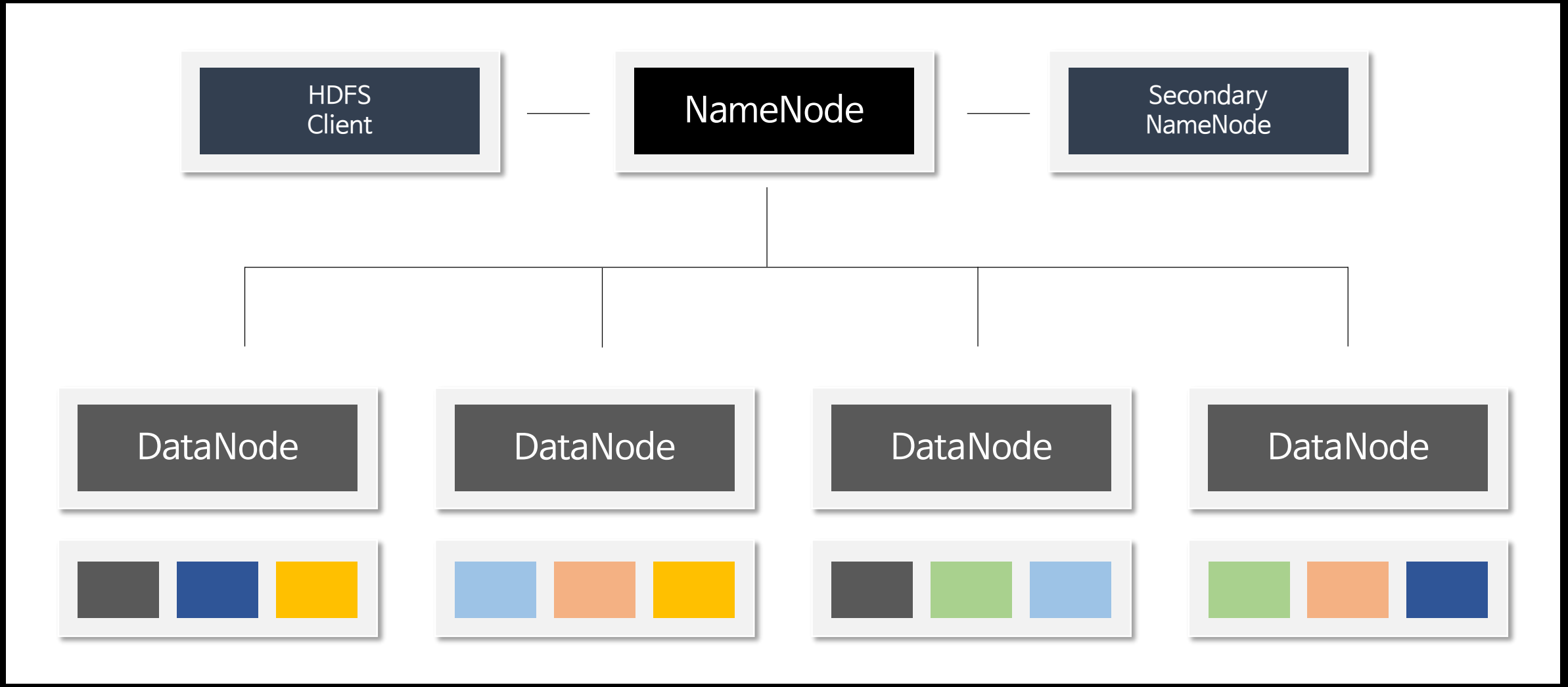
HDFS (Hadoop Distribution File System)?

- HDFS란 무엇일까? -

"Hadoop 클러스터의 데이터를
저장하는 분산형 파일 시스템"

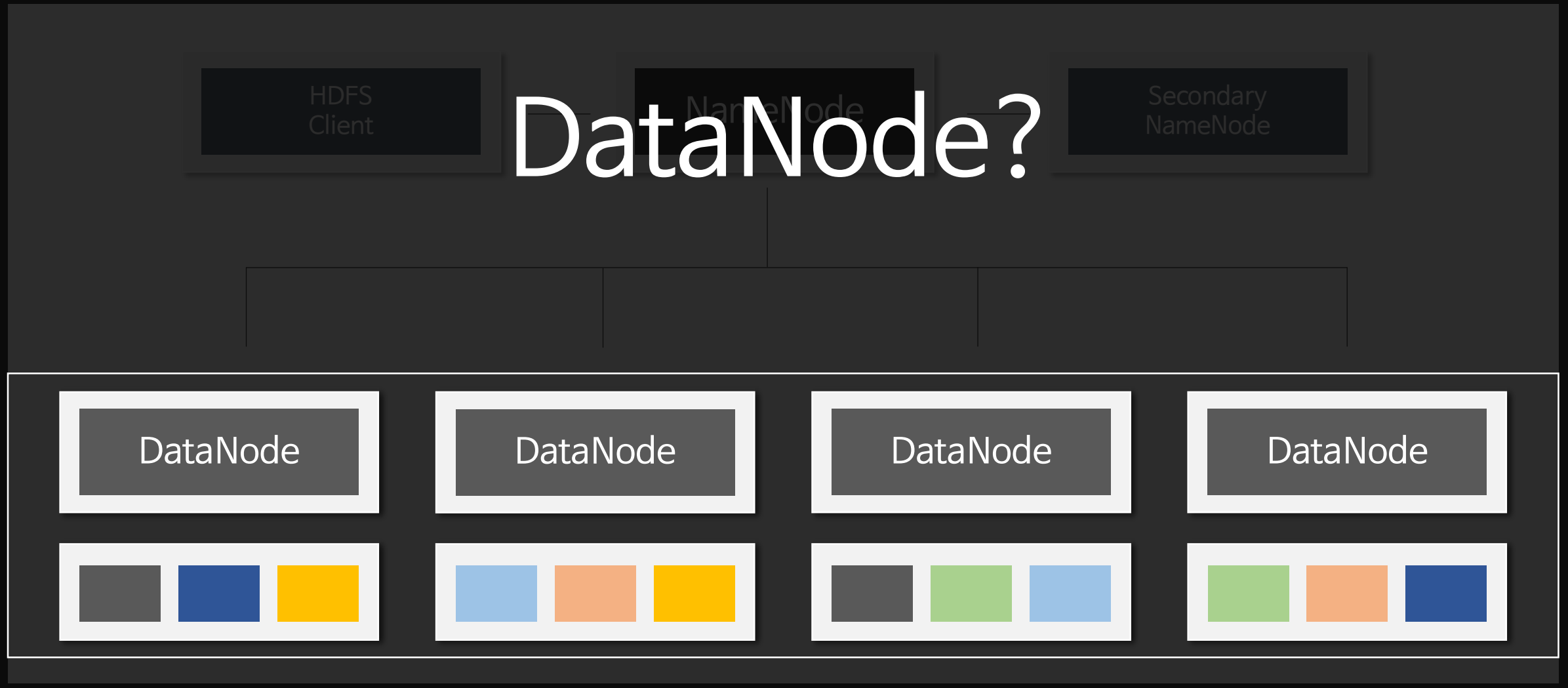
HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -



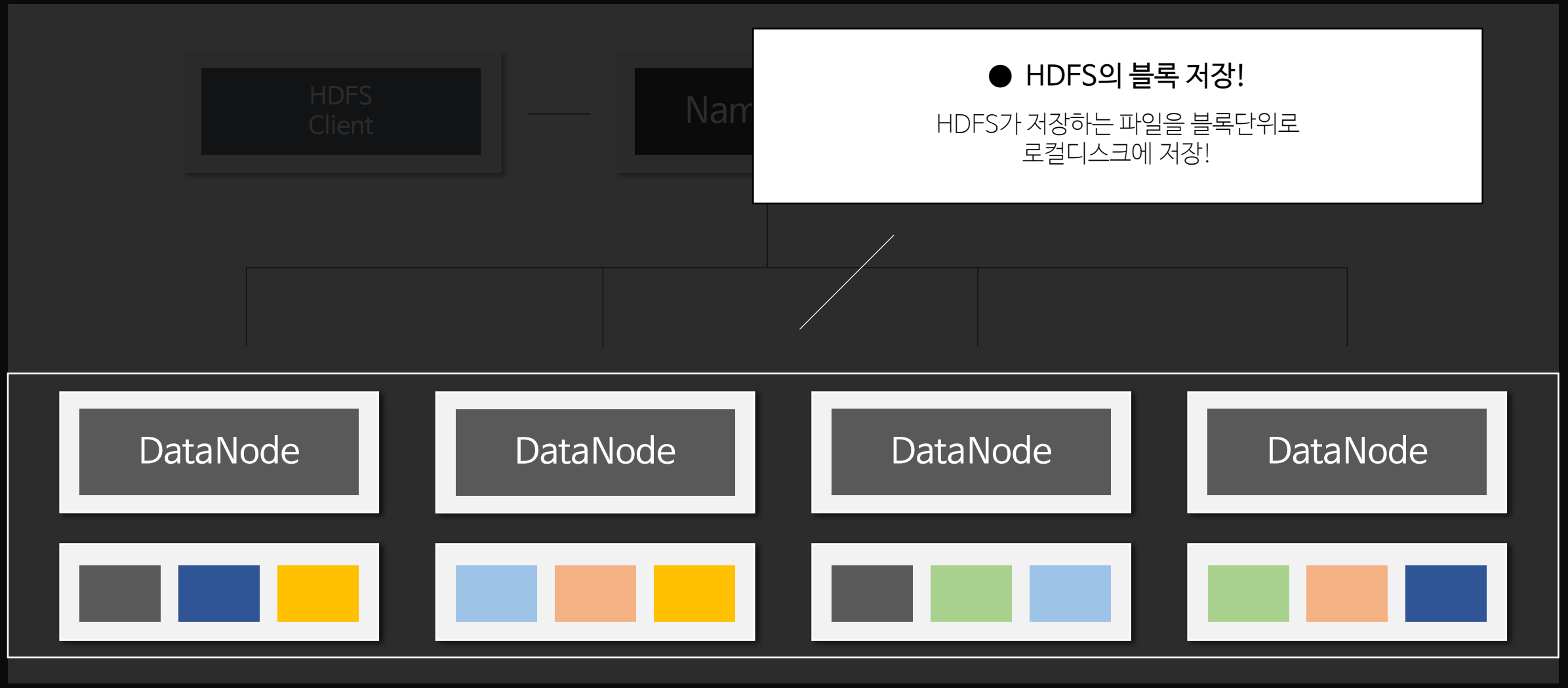
HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -



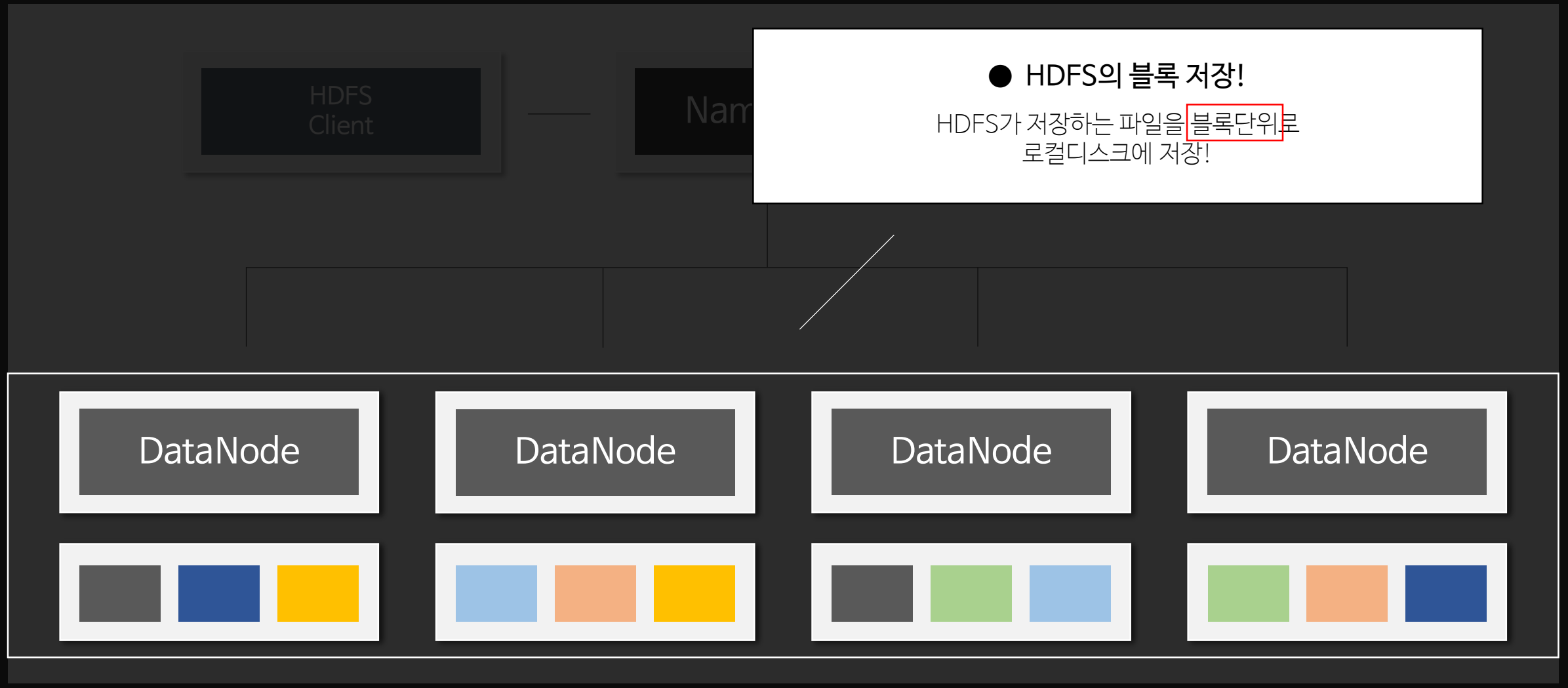
HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -



HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -

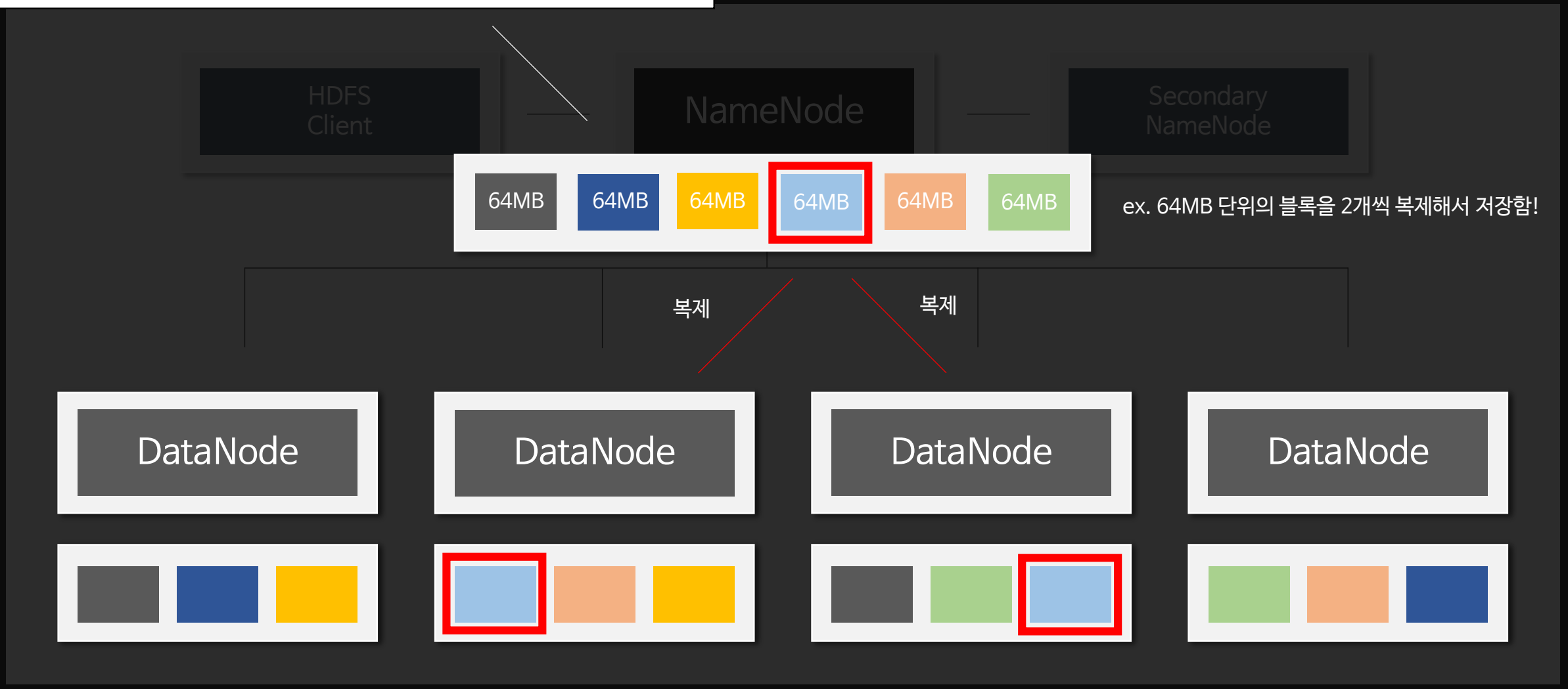


stribution File System)?

분산 파일 시스템 –

● **블록구조 파일 시스템**

파일을 블록 단위로 나눠서 분산 저장함
데이터 블록사이즈와 복제 개수는 설정해주어야 함!

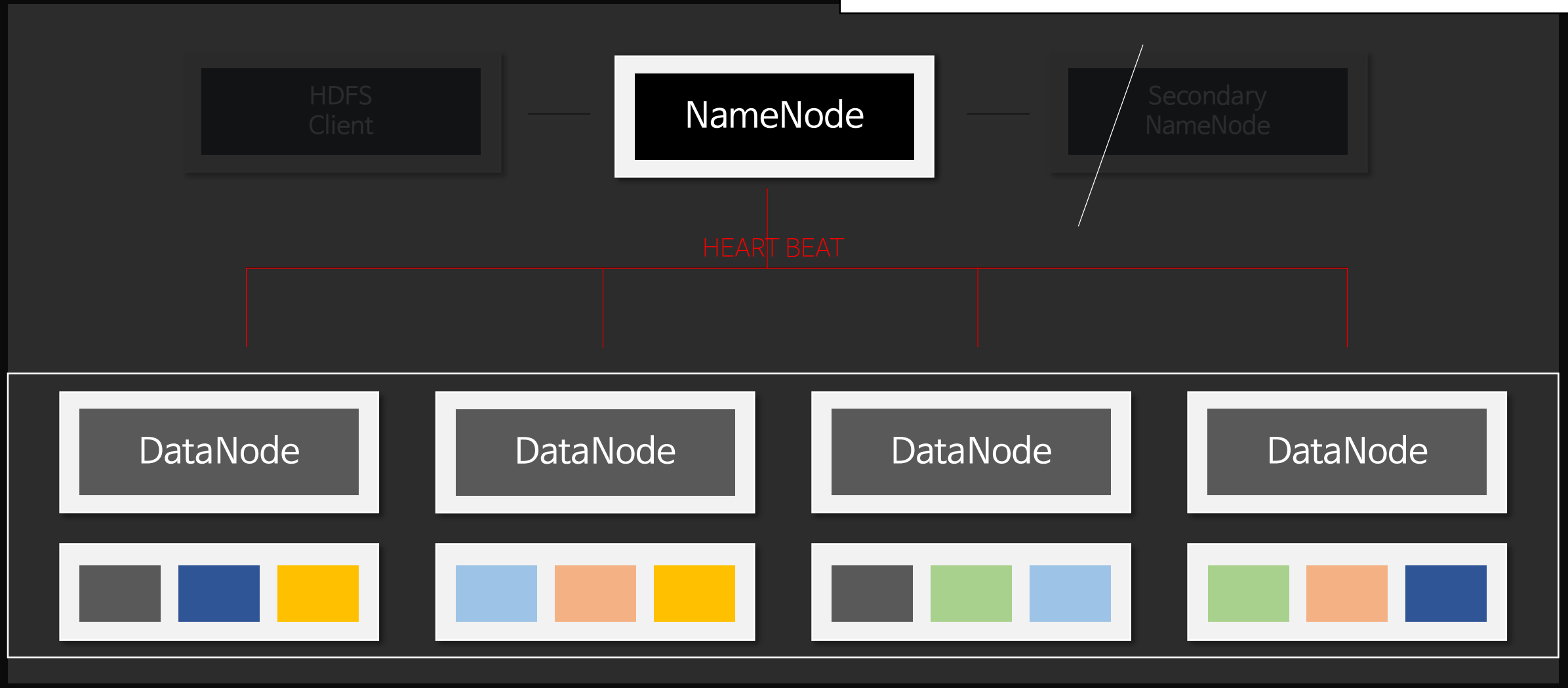


HDFS (Hadoop Distribu

- 하둡 분산 파일

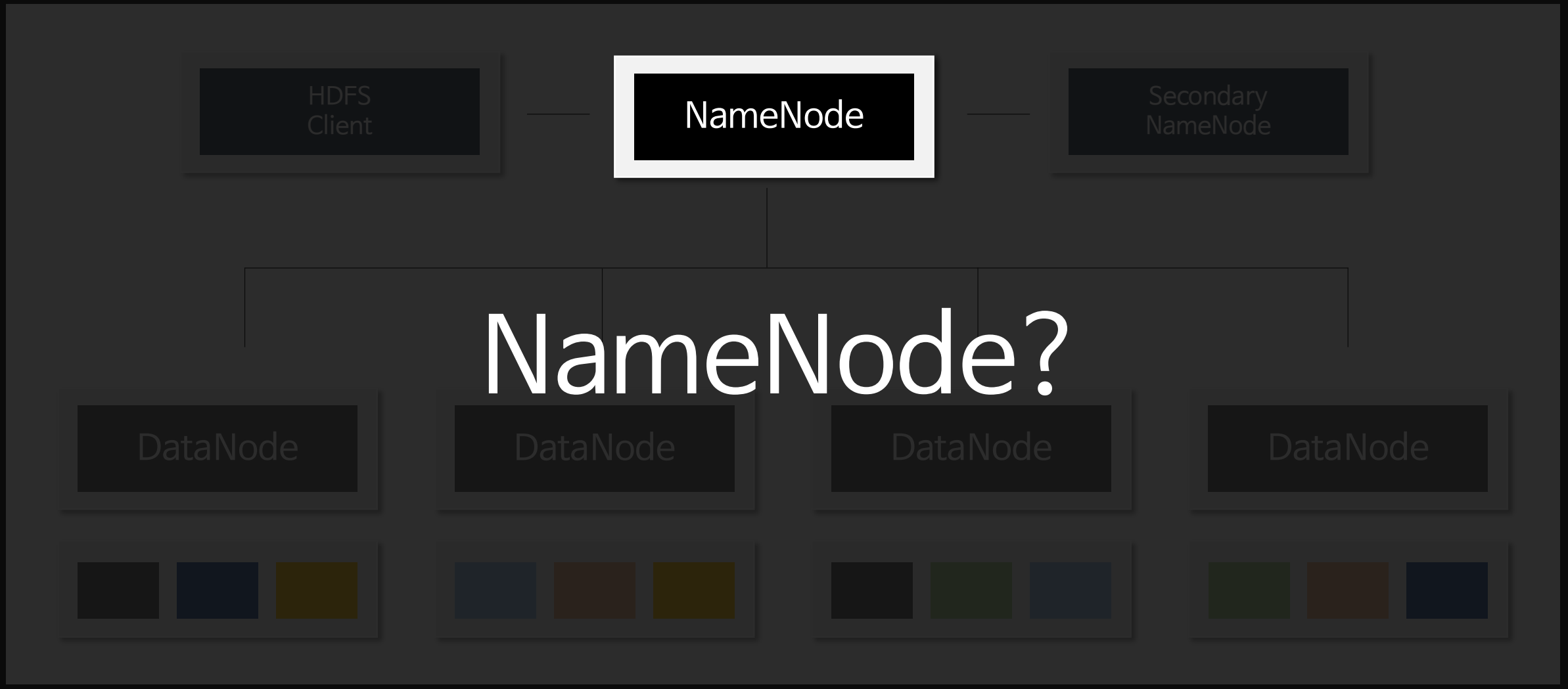
● 네임노드에 하트비트 전송!

3초마다 한번씩 네임노드에게
하트비트를 전송함으로써 상태를 보고함



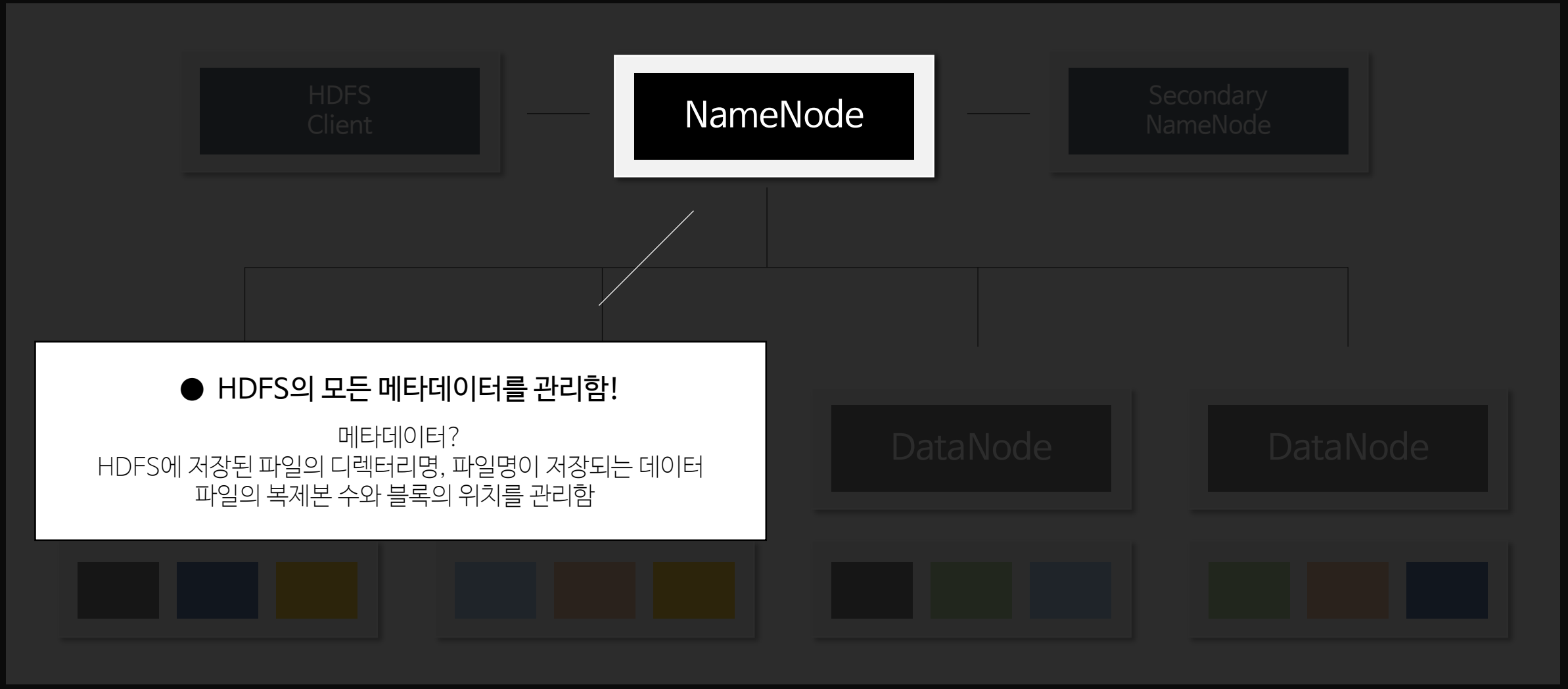
HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -



HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -

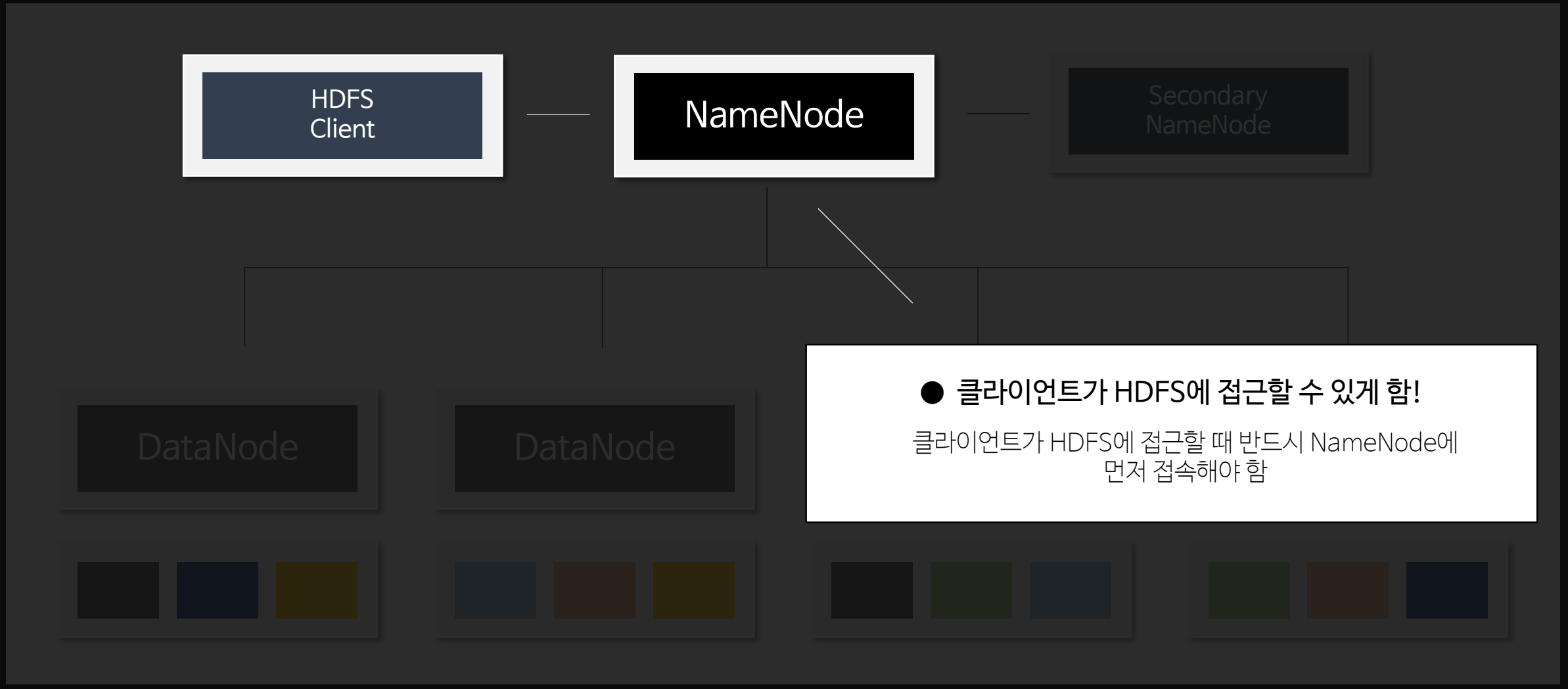


● **HDFS의 모든 메타데이터를 관리함!**

메타데이터?
HDFS에 저장된 파일의 디렉터리명, 파일명이 저장되는 데이터
파일의 복제본 수와 블록의 위치를 관리함

HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -

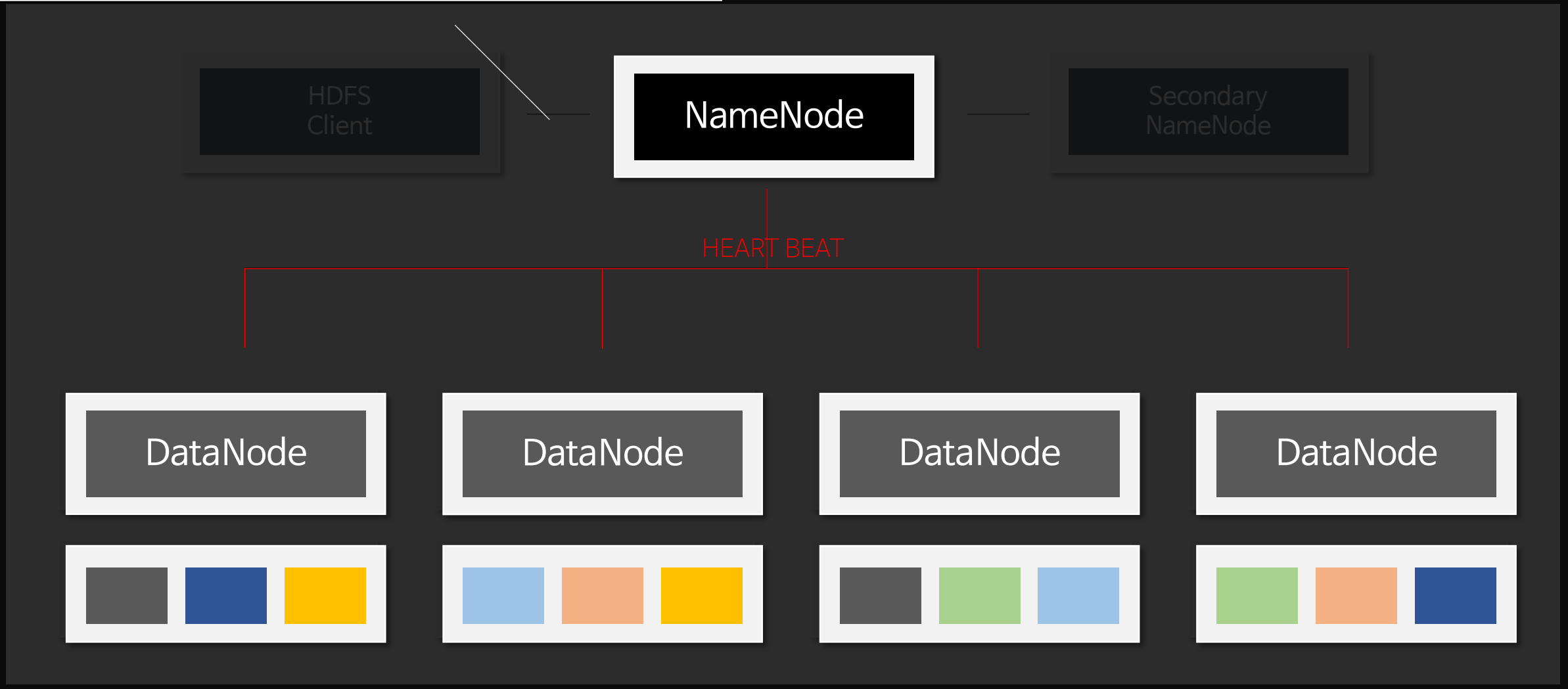


stribution File System)?

– 분산 파일 시스템 –

● **데이터노드 모니터링!**

주기적으로 하트비트를 전송하지 않는 데이터노드는
장애서버로 판단

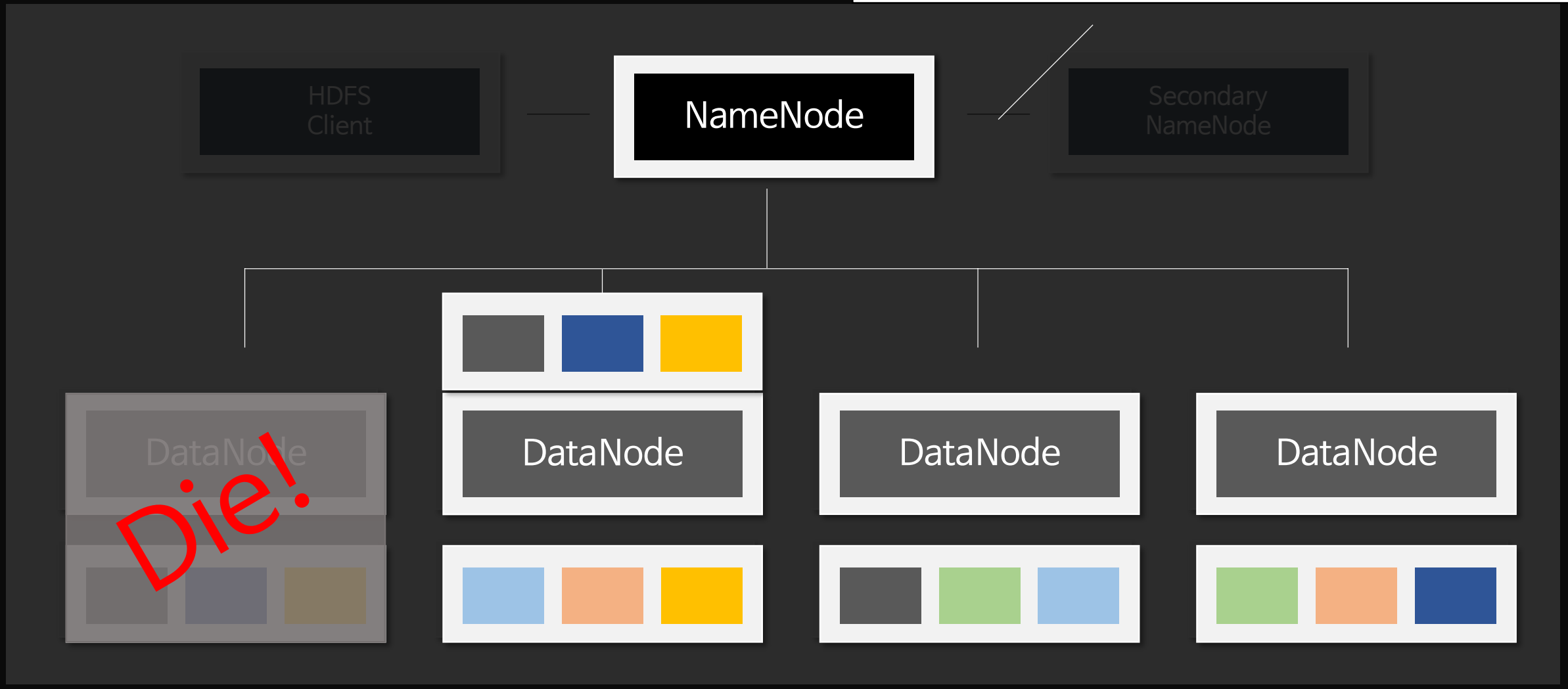


HDFS (Hadoop Distribut

- 하둡 분산 파일 시

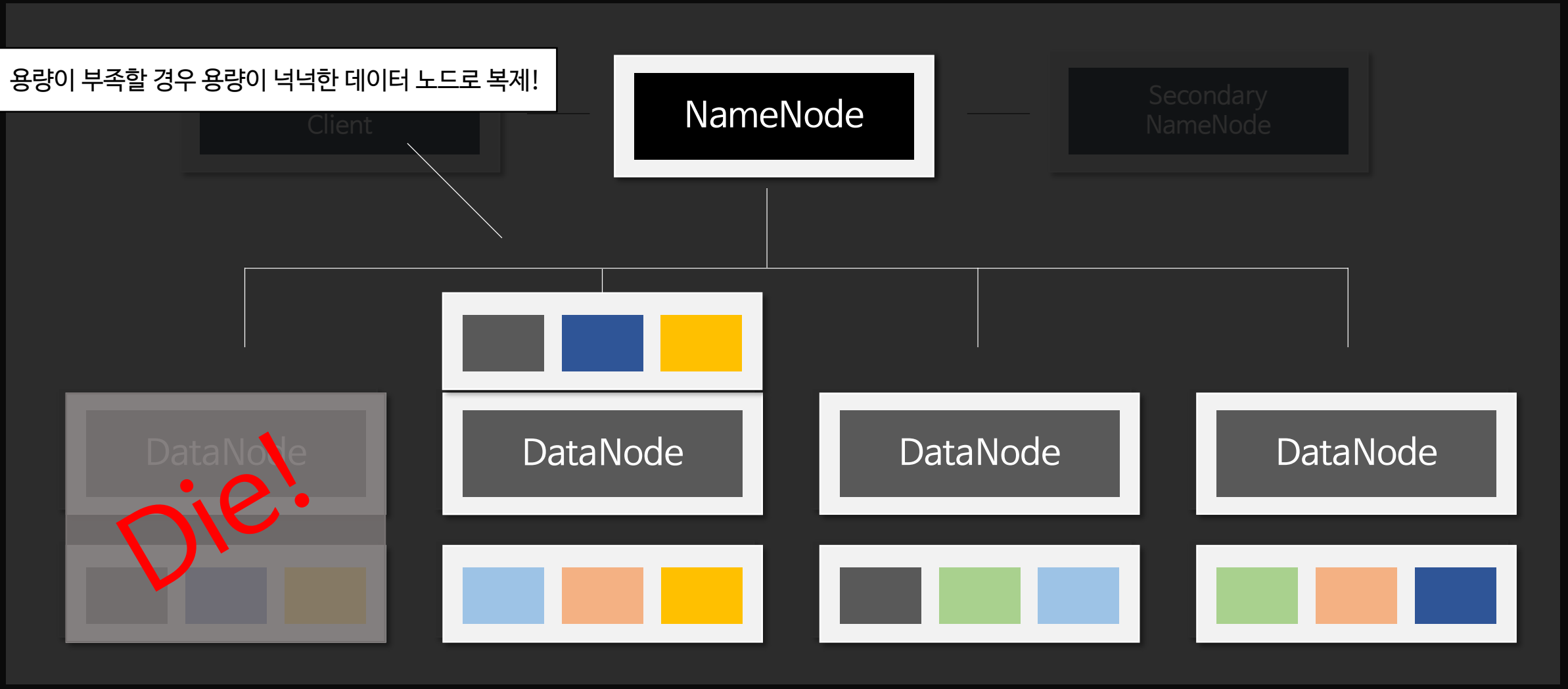
● **블록 관리**

장애가 발생한 데이터노드의 블록을
다른 데이터노드로 복제



HDFS (Hadoop Distribution File System)?

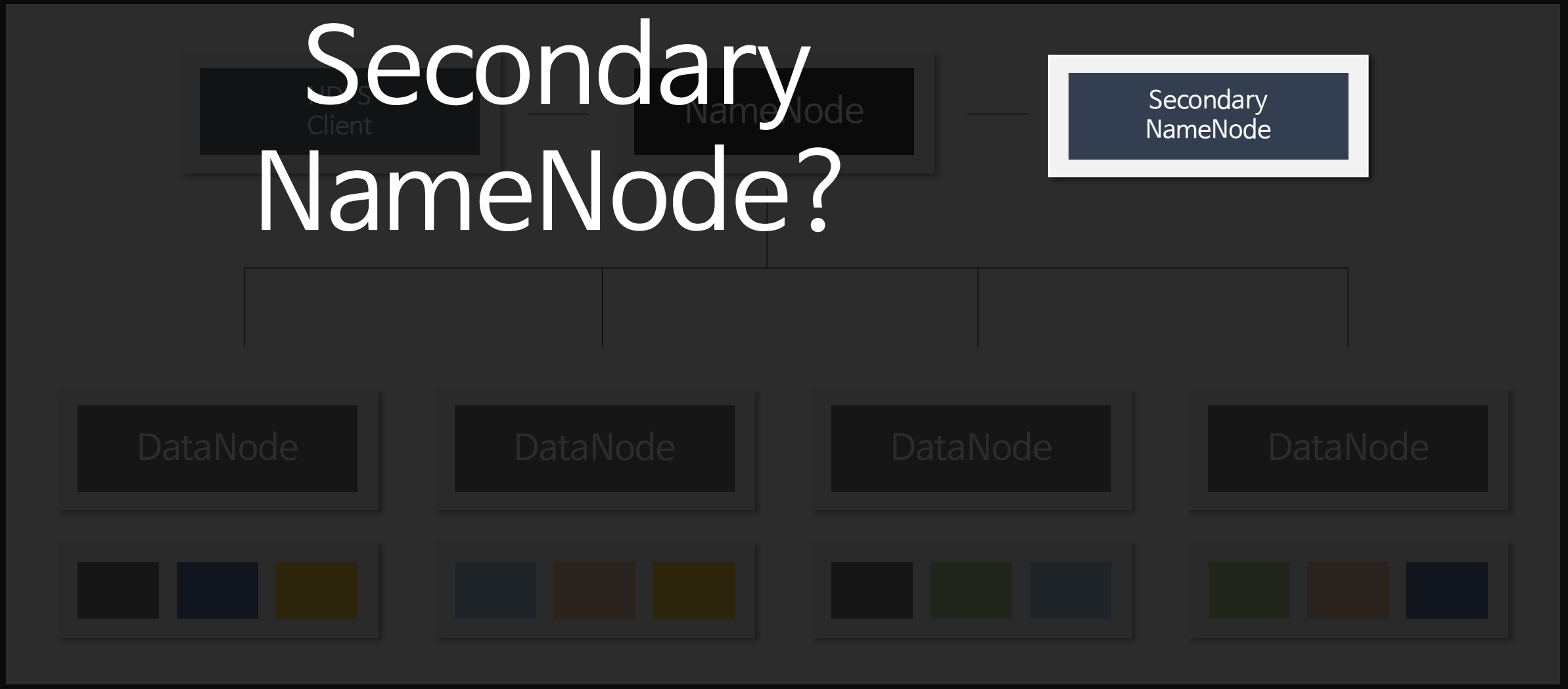
- 하둡 분산 파일 시스템 -



HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -

Secondary NameNode?



HDFS (Hadoop Distribution File System)?

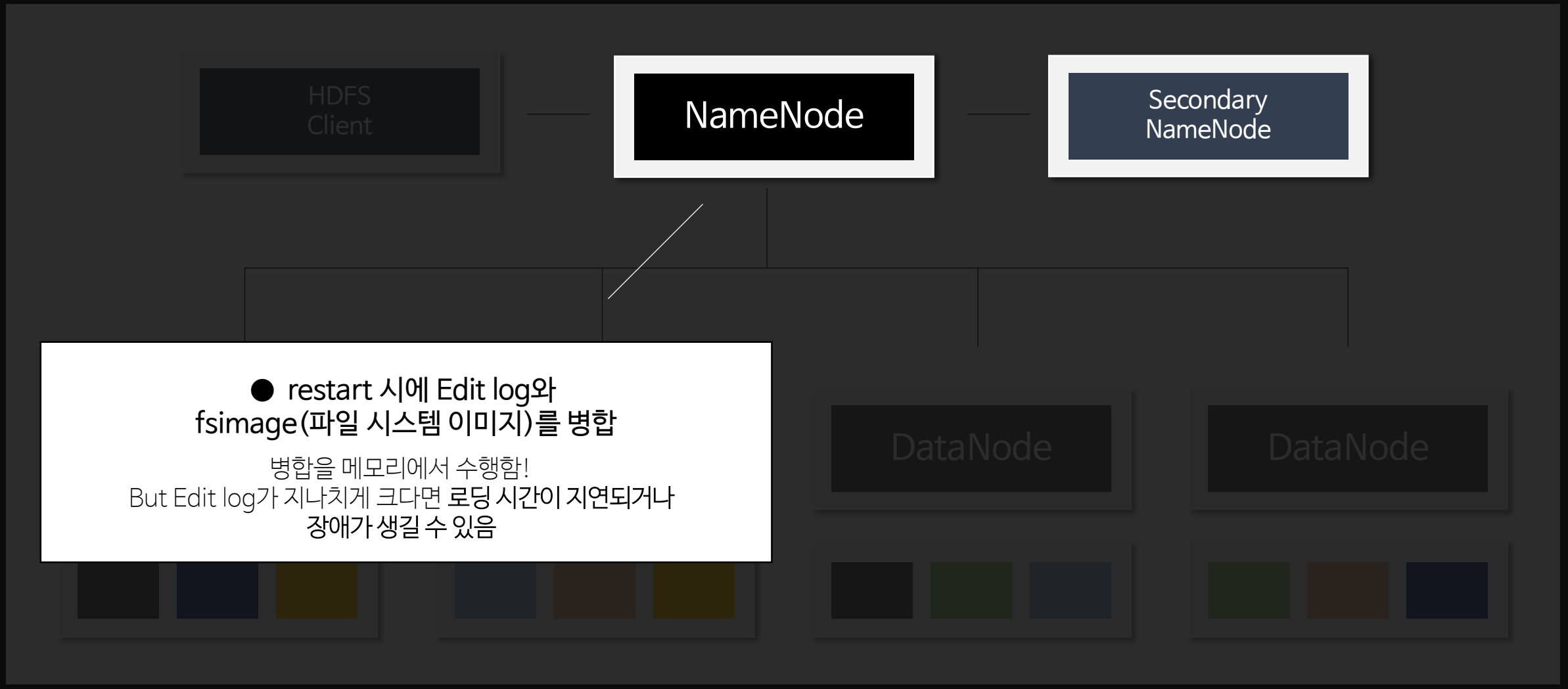
- 하둡 분산 파일 시스템 -

- NameNode는 HDFS내에서 발생하는 트랜잭션을 Edit log(에디트 로그)에 저장함!

  로그 파일에 용량 제한이 없어서 무한하게 저장 가능!

HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -

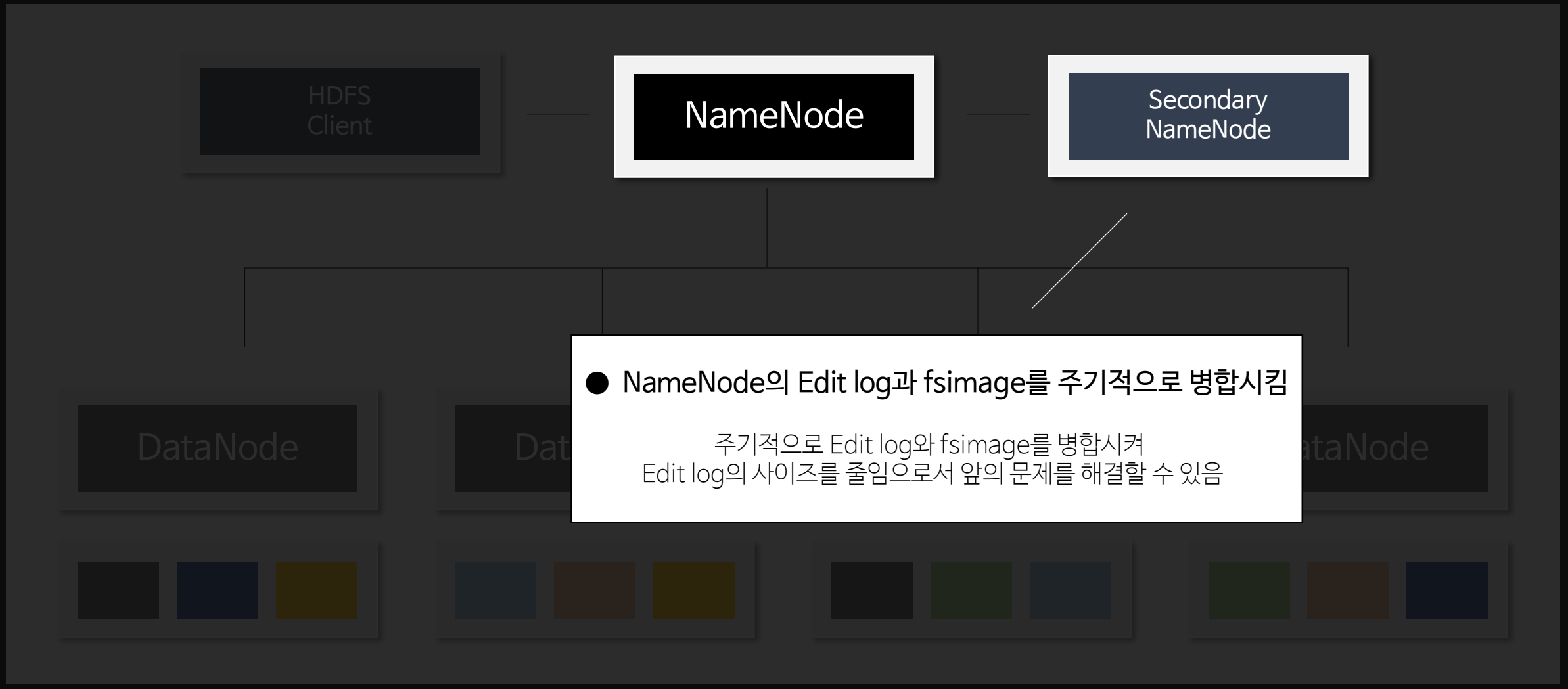


● restart 시에 Edit log와
fsimage(파일 시스템 이미지)를 병합

병합을 메모리에서 수행함!
But Edit log가 지나치게 크다면 **로딩 시간이 지연되거나**
장애가 생길 수 있음

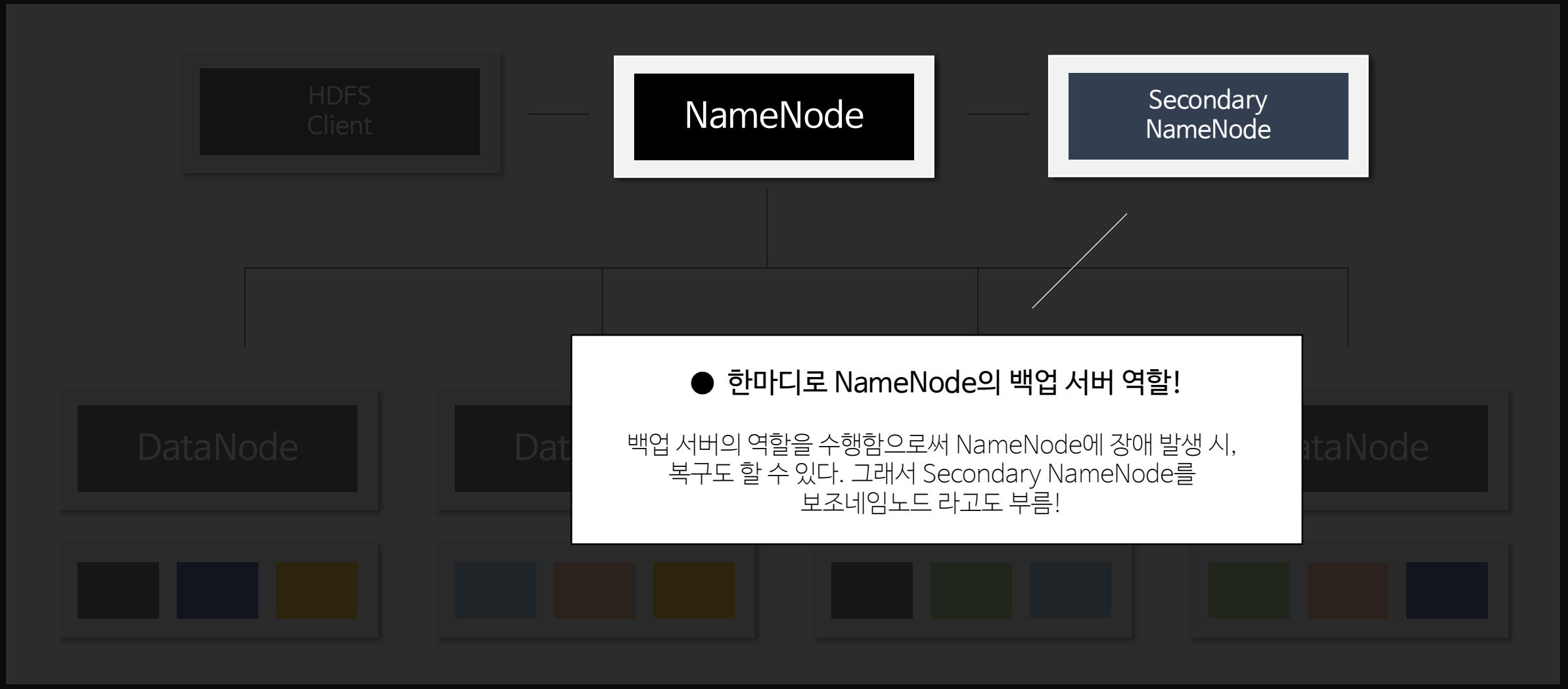
HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -



HDFS (Hadoop Distribution File System)?

- 하둡 분산 파일 시스템 -



Hadoop?

- 하둡이란 무엇일까? -

HDFS

분산 파일 **저장** 시스템

MapReduce

분산 파일 **처리** 시스템

MapReduce?

"대용량의 데이터를 **HDFS**에 저장하고,
MapReduce로 처리하여 분산처리 한다!"

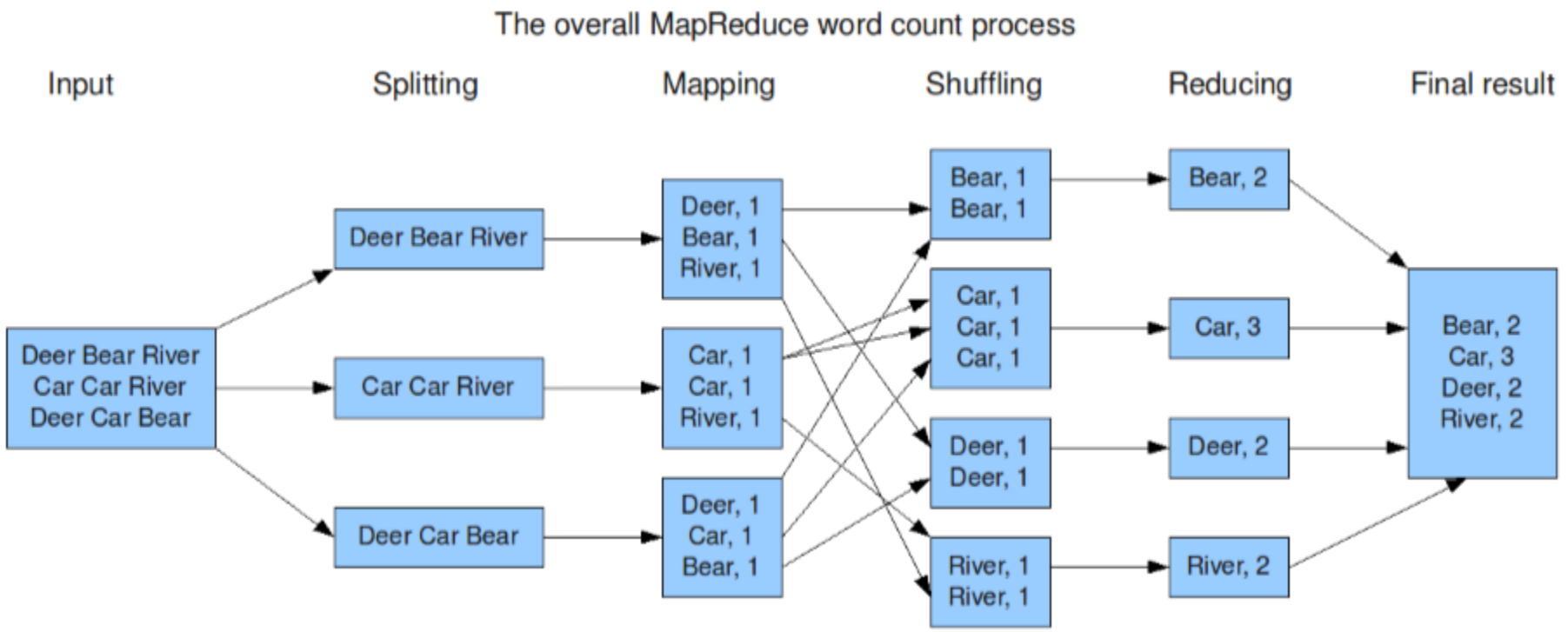
MapReduce?

- MapReduce이란 무엇일까? -

"Hadoop 클러스터의 데이터를
처리하기 위한 시스템"

MapReduce?

- 맵리듀스 처리 흐름 -



"Map단계와 Reduce단계를 거쳐서 처리한다!"

MapReduce?

- 맵리듀스 처리 흐름-

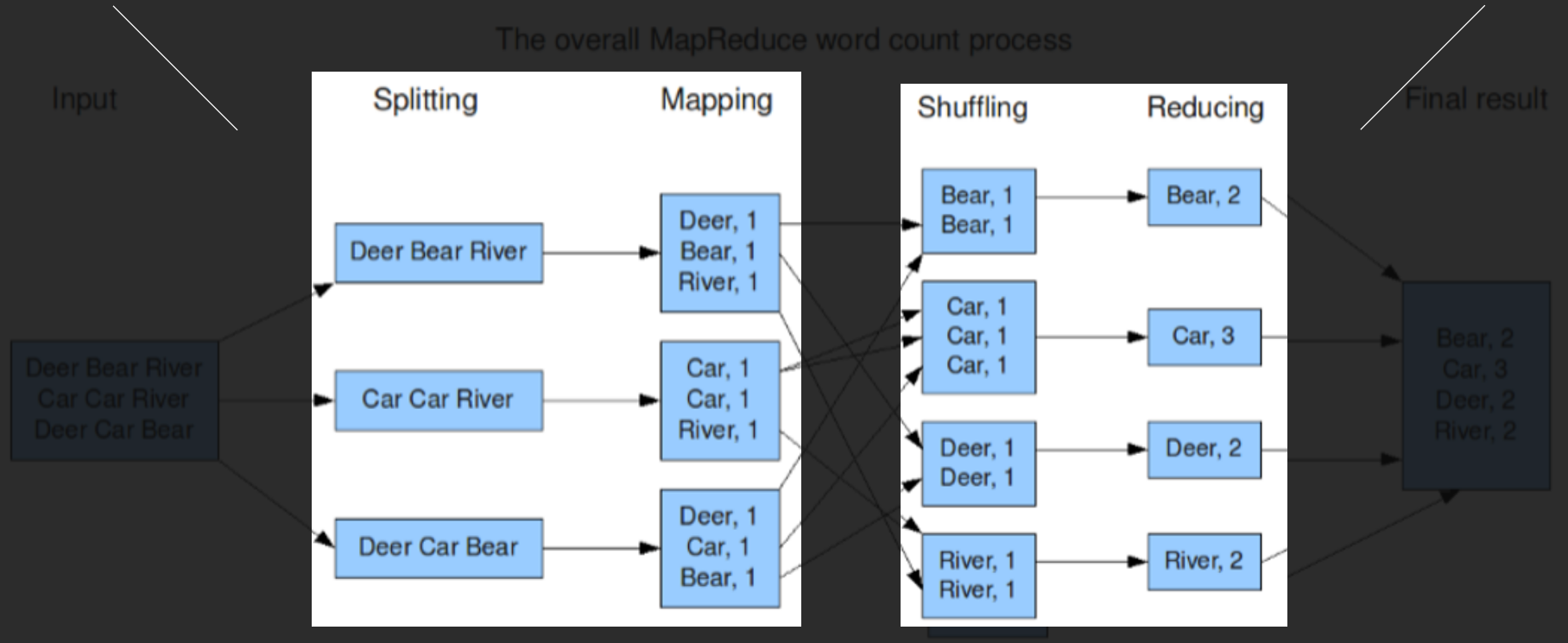
● Map

Split한 데이터를 (Key, Value) 형태로
연관성 있는 데이터로 묶는 작업

**기본적으로 하나의 HDFS 블록을
대상으로 수행함!**

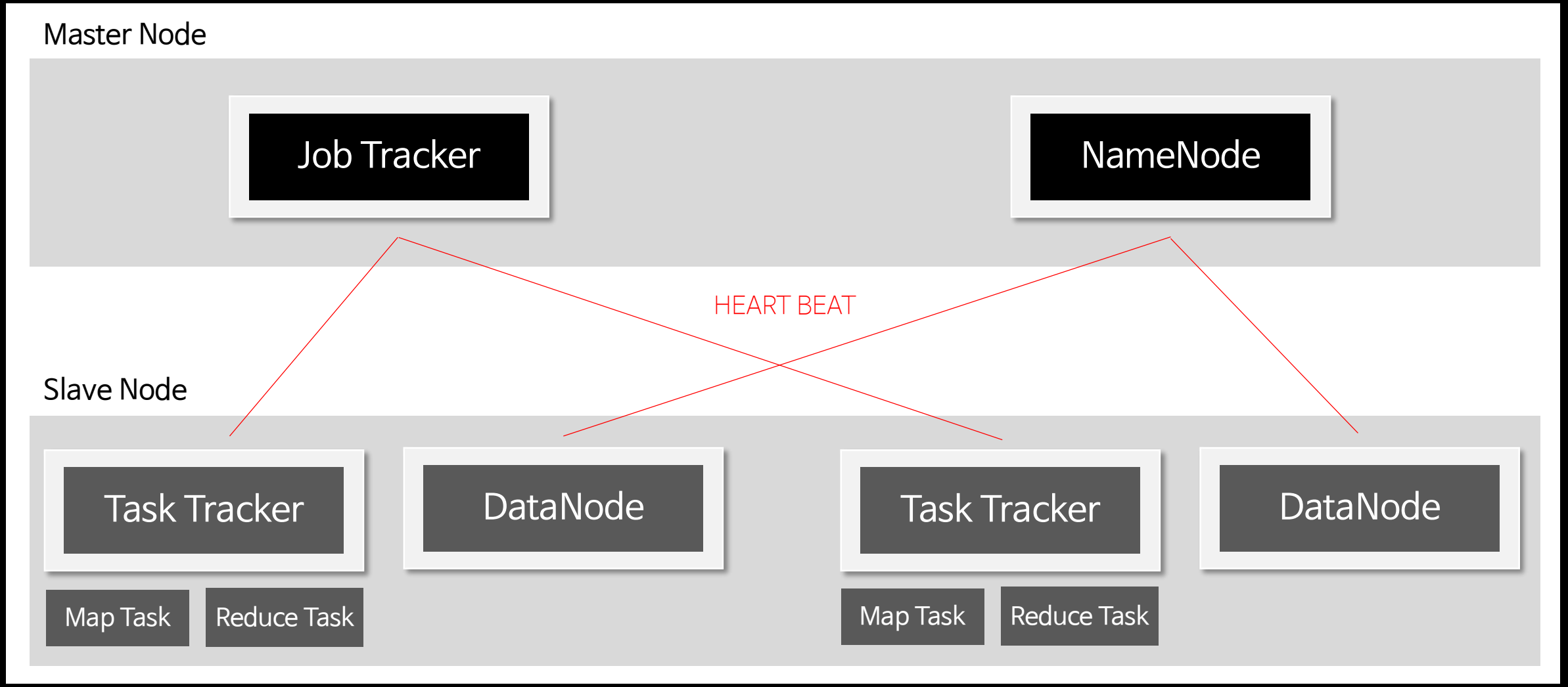
● Reduce

Map한 결과에서 중복 있는 데이터를
제거하고, 원하는 데이터 추출



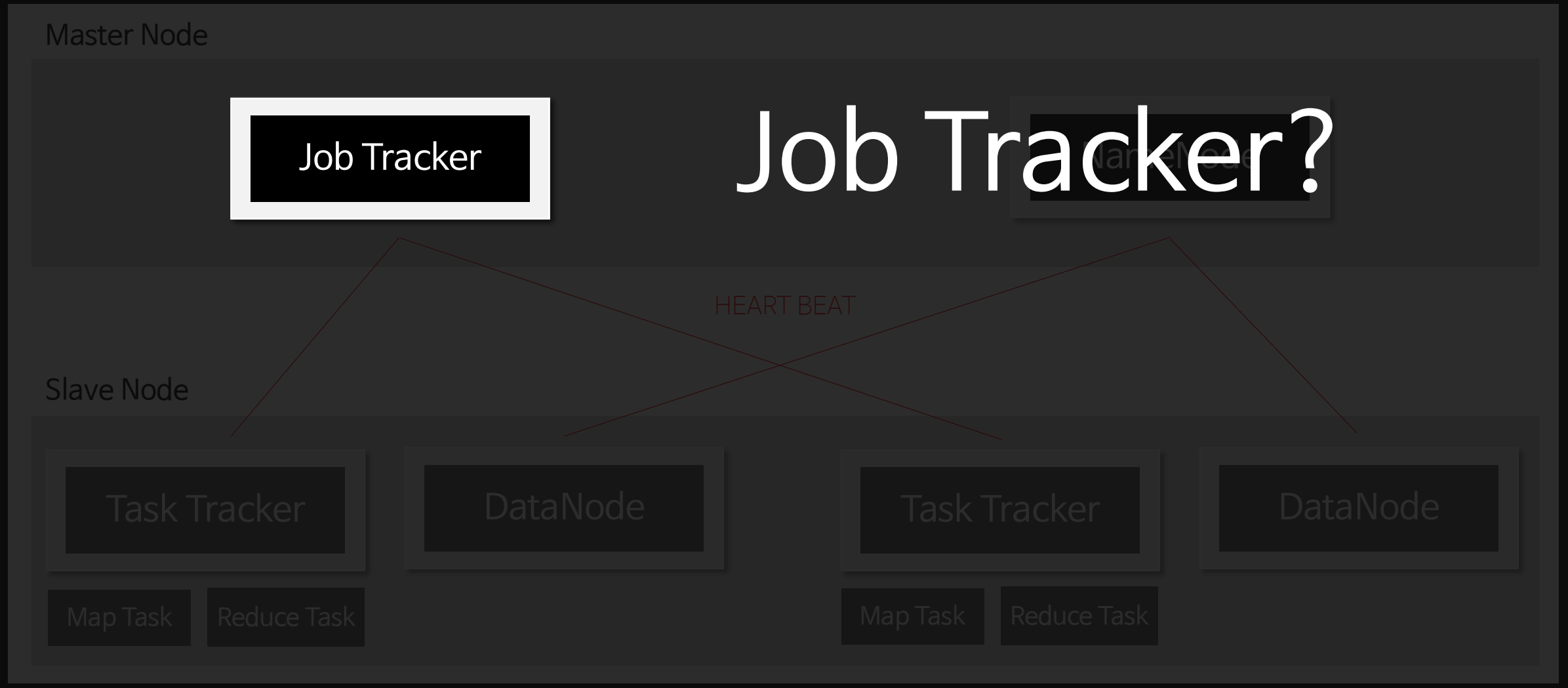
MapReduce?

- 맵리듀스 작동 방식-



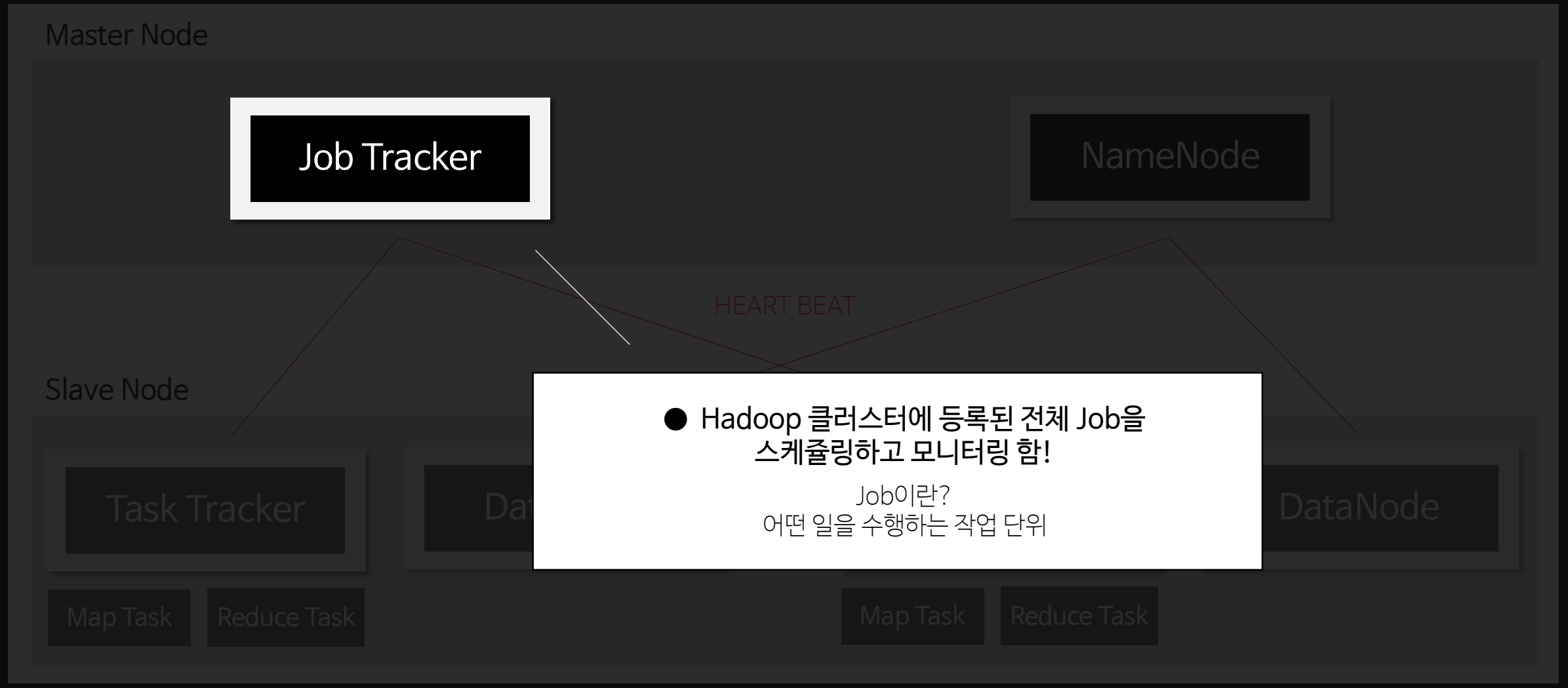
MapReduce?

- 맵리듀스 작동 방법-



MapReduce?

- 맵리듀스 작동 방식 -

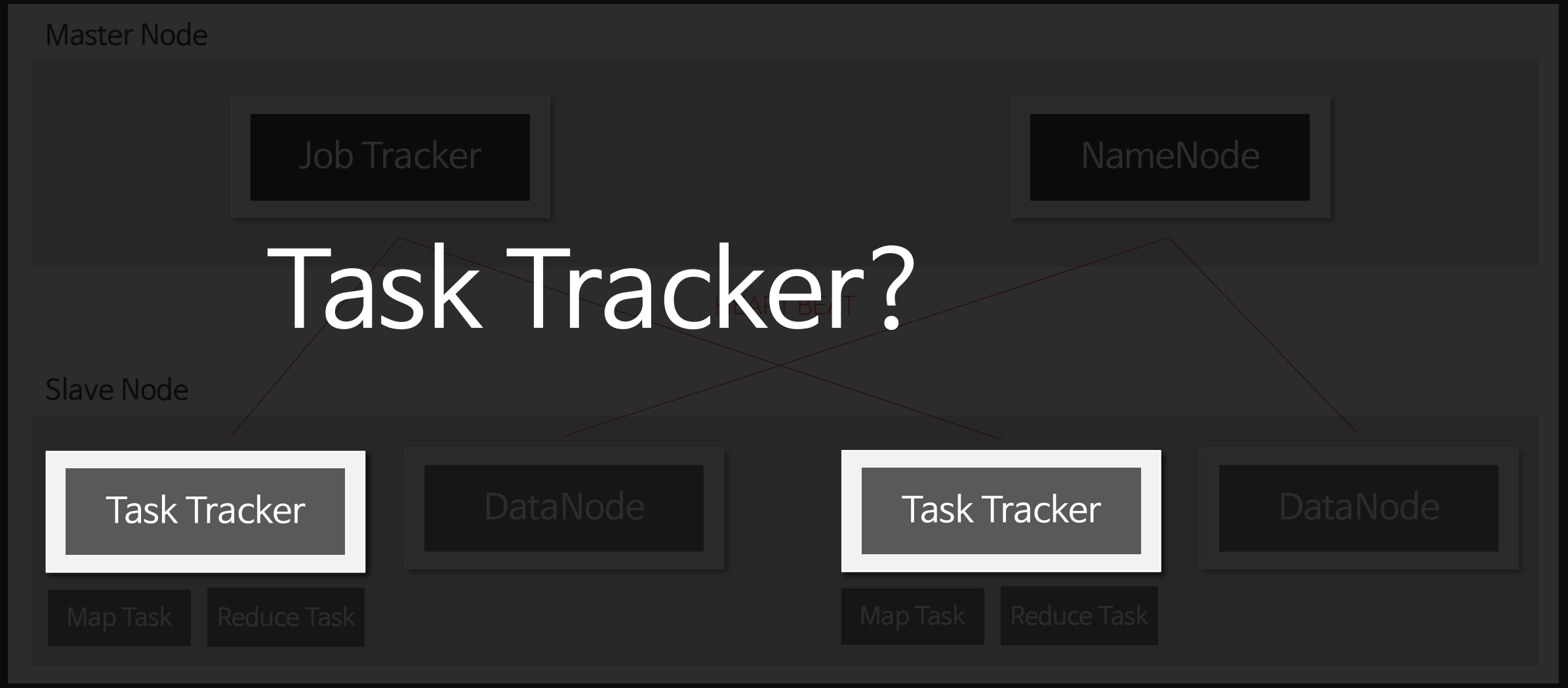


- Hadoop 클러스터에 등록된 전체 Job을 스케쥴링하고 모니터링 함!

Job이란?
어떤 일을 수행하는 작업 단위

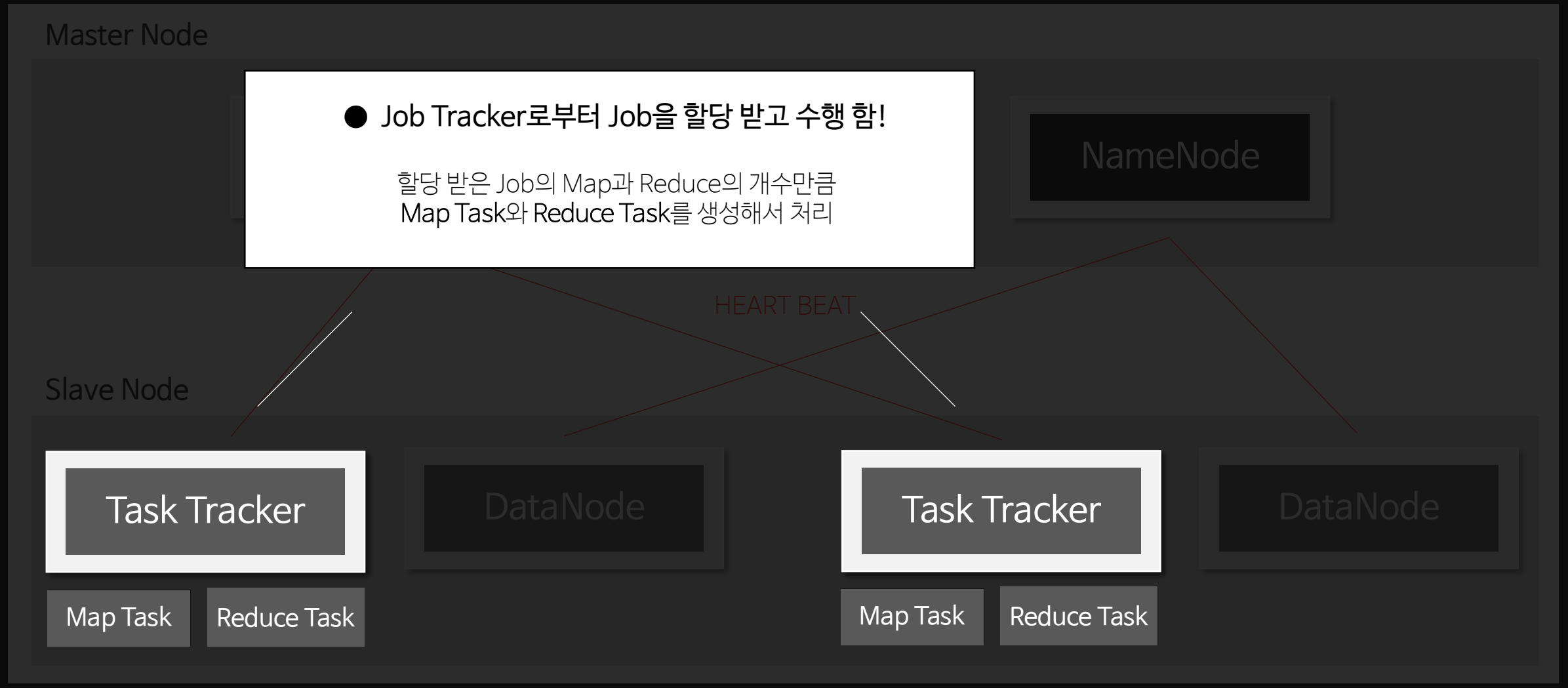
MapReduce?

- 맵리듀스 작동 방법-



MapReduce?

– 맵리듀스 작동 방식 –



과제 내용

Condition A

HDFS에 저장되는 파일을 블록단위로
저장하는 DataNode를 3개 이상
&
이를 관리하는 NameNode
로 구성된 HDFS 를 만들어라!

● 3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것

A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
C. Data Block Size는 32MB로 설정
D. Block Replication(복제)는 2로 설정
E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

● Hive 서비스 제공

A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
Hive JDBC를 제공할 것
데이터가 존재하는 테이블을 3개 이상 만들 것

서비스 제공

lin notebook 서비스 제공

과제 내용

Condition A

- 3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것
  - A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  - B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  - C. Data Block Size는 32MB로 설정
  - D. Block Replication(복제)는 2로 설정
  - E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공
  - A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  - Hive JDBC를 제공할 것
  - 데이터가 존재하는 테이블을 3개 이상 만들 것

- 서비스 제공

- Zeppelin notebook 서비스 제공

이 때, 저장되는 블록의 사이즈는 32MB!
복제하는 블록의 개수는 2개!
로 설정해라!

과제 내용

Condition A

?

- 3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것

  A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  C. Data Block Size는 32MB로 설정
  D. Block Replication(복제)는 2로 설정
  E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공

  A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  Hive JDBC를 제공할 것
  데이터가 존재하는 테이블을 3개 이상 만들 것

서비스 제공

- Zeppelin notebook 서비스 제공

과제 내용

Condition A

"하둡 에코시스템"에 대해 알면 된다!

- 3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것

  A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  C. Data Block Size는 32MB로 설정
  D. Block Replication(복제)는 2로 설정
  E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공

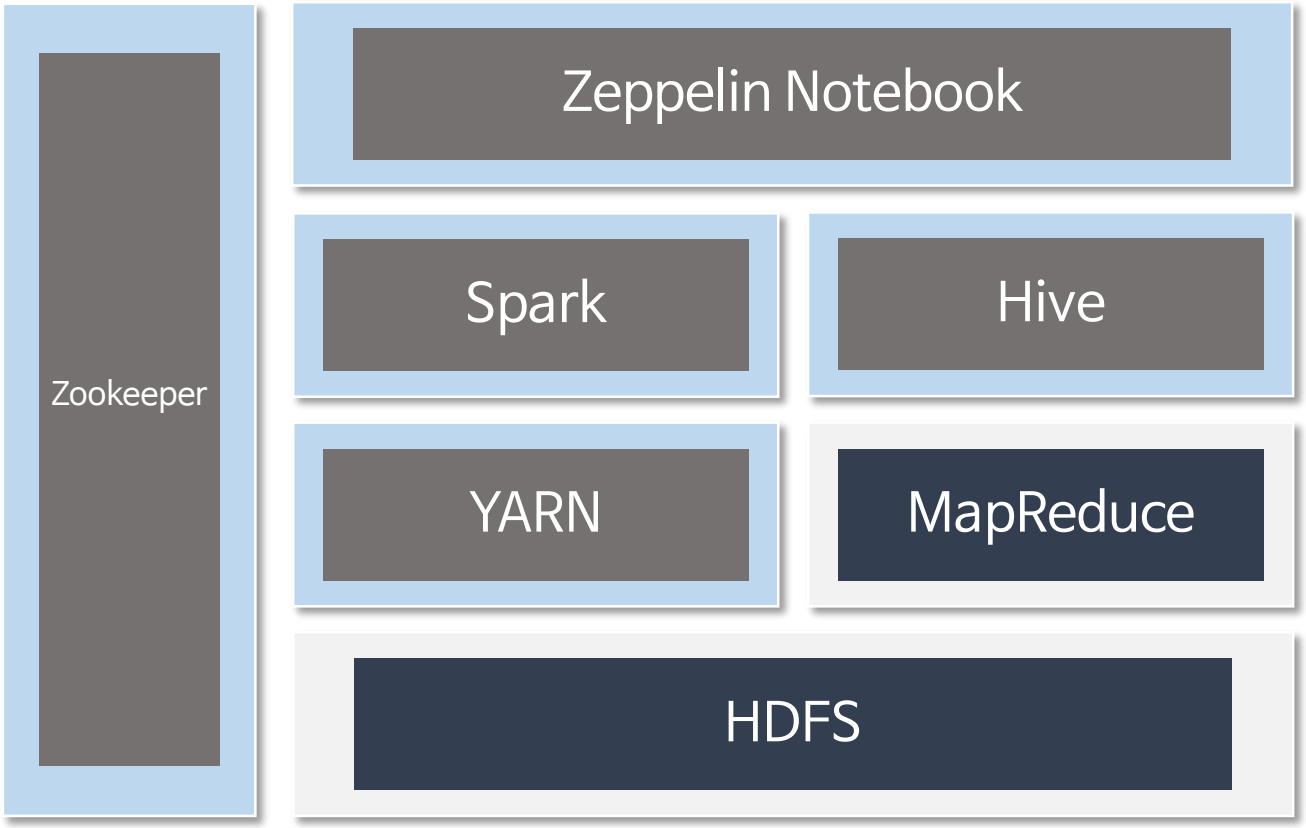
  A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  Hive JDBC를 제공할 것
  데이터가 존재하는 테이블을 3개 이상 만들 것

서비스 제공

- Zeppelin notebook 서비스 제공

Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -

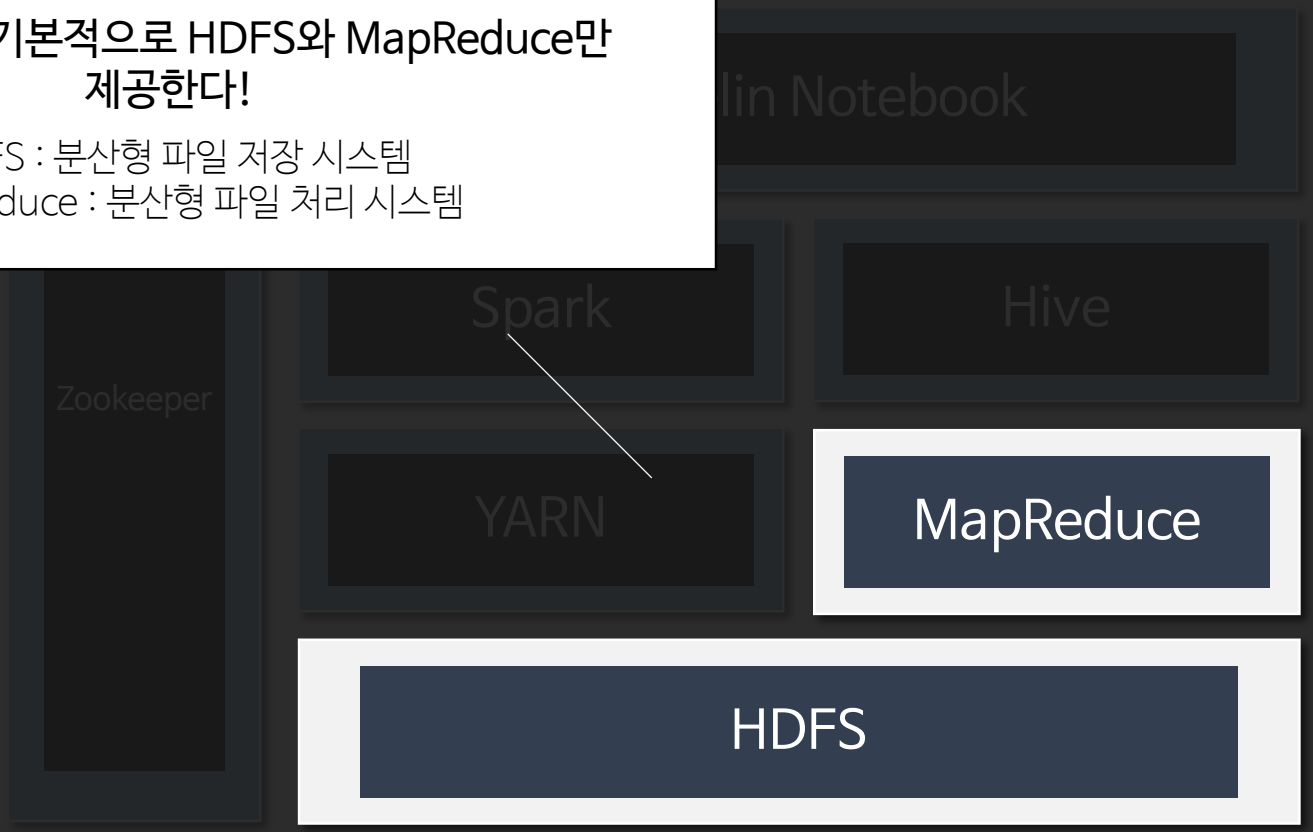


Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -

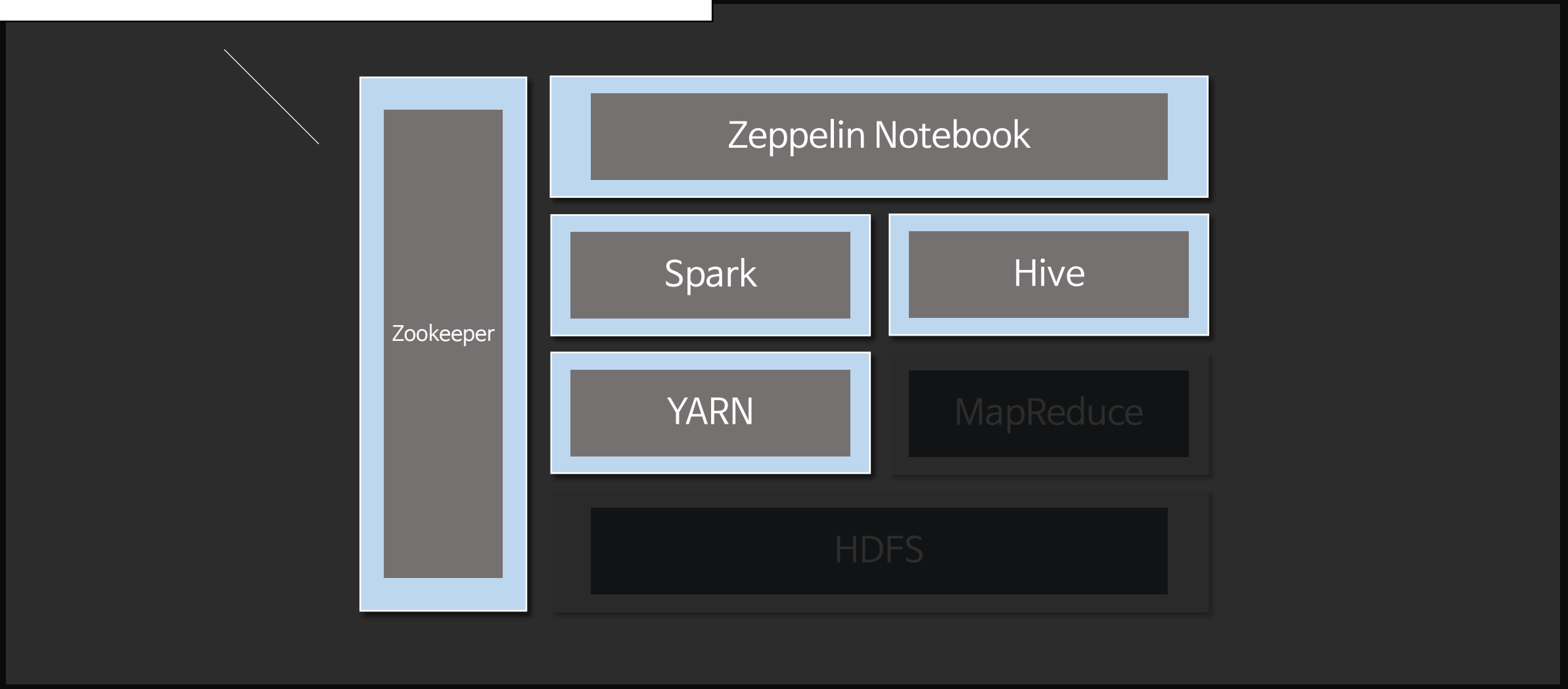
● Hadoop은 기본적으로 HDFS와 MapReduce만 제공한다!

HDFS : 분산형 파일 저장 시스템
MapReduce : 분산형 파일 처리 시스템



● HDFS, MapReduce 제외 서브 프로젝트(어플리케이션)들을 사용 할 수 있음!

기존의 HDFS와 MapReduce의 단점을 보완하고,
좀 더 효율적으로 적용할 수 있도록 서브 프로젝트들을 같이 사용함

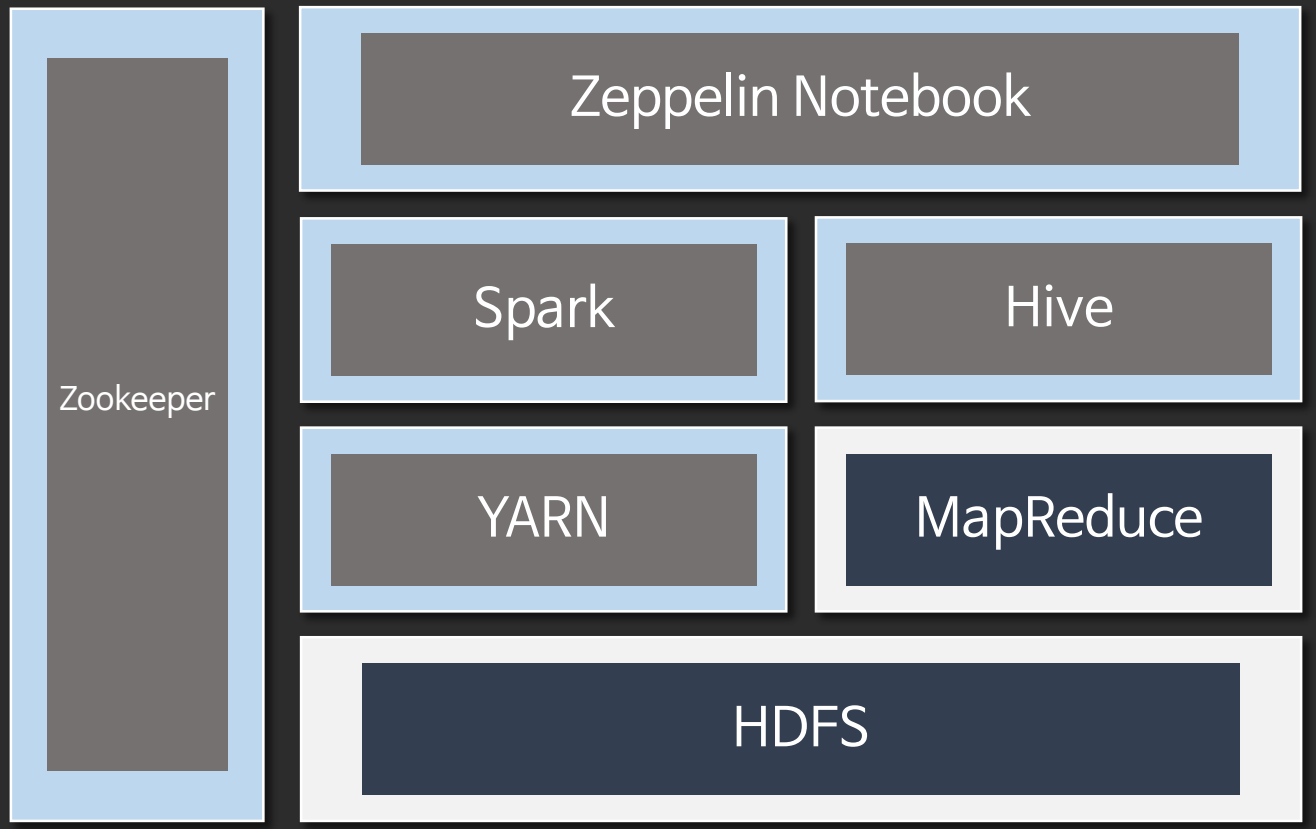


Hadoop eco

– 하둡 에코시스템이

● **하둡 에코시스템!**

HDFS / MapReduce 와 기타 서브 프로젝트들을
통틀어 하둡 에코시스템이라고 함!



과제 내용

Condition A

YARN / Zookeeper를 클러스터에 깔아라!

- 3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것
  - A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  - B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  - C. Data Block Size는 32MB로 설정
  - D. Block Replication(복제)는 2로 설정
  - E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공
  - A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  - Hive JDBC를 제공할 것
  - 데이터가 존재하는 테이블을 3개 이상 만들 것

- 서비스 제공

- Zeppelin notebook 서비스 제공

Hive를 클러스터에
깔아라!

과제 내용

Condition A

3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것

A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
C. Data Block Size는 32MB로 설정
D. Block Replication(복제)는 2로 설정
E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

● Hive 서비스 제공

A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
B. Hive JDBC를 제공할 것
C. 데이터가 존재하는 테이블을 3개 이상 만들 것

● Spark 서비스 제공

● Zeppelin notebook 서비스 제공

Spark / Zeppelin notebook을 클러스터에 깔아라!

Condition A

- 3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것
  - A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  - B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  - C. Data Block Size는 32MB로 설정
  - Block Replication(복제)는 2로 설정
  - NameNode와 Hive service는 각각 다른 HOST에서 구동할 것

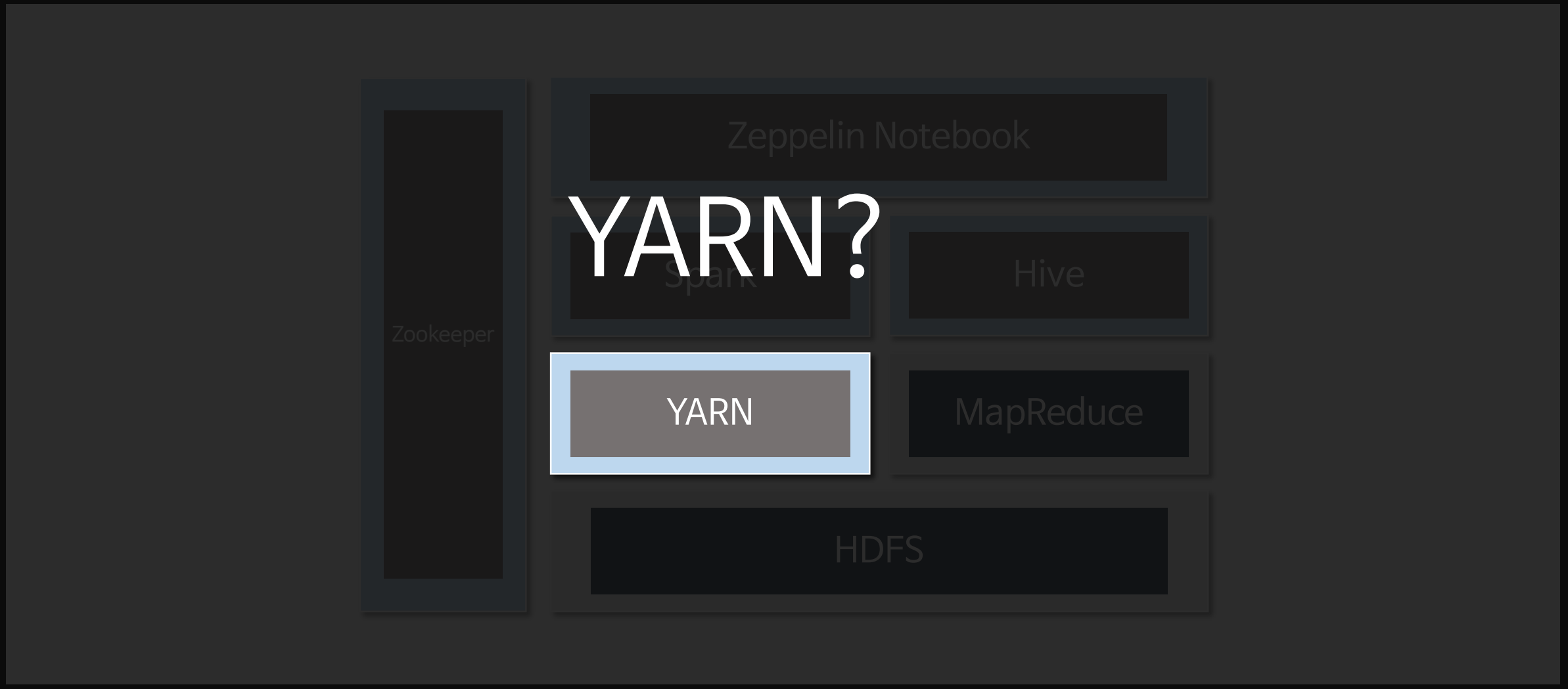
서비스 제공

- A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
- B. Hive JDBC를 제공할 것
- C. 데이터가 존재하는 테이블을 3개 이상 만들 것

- Spark 서비스 제공
- Zeppelin notebook 서비스 제공

Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -



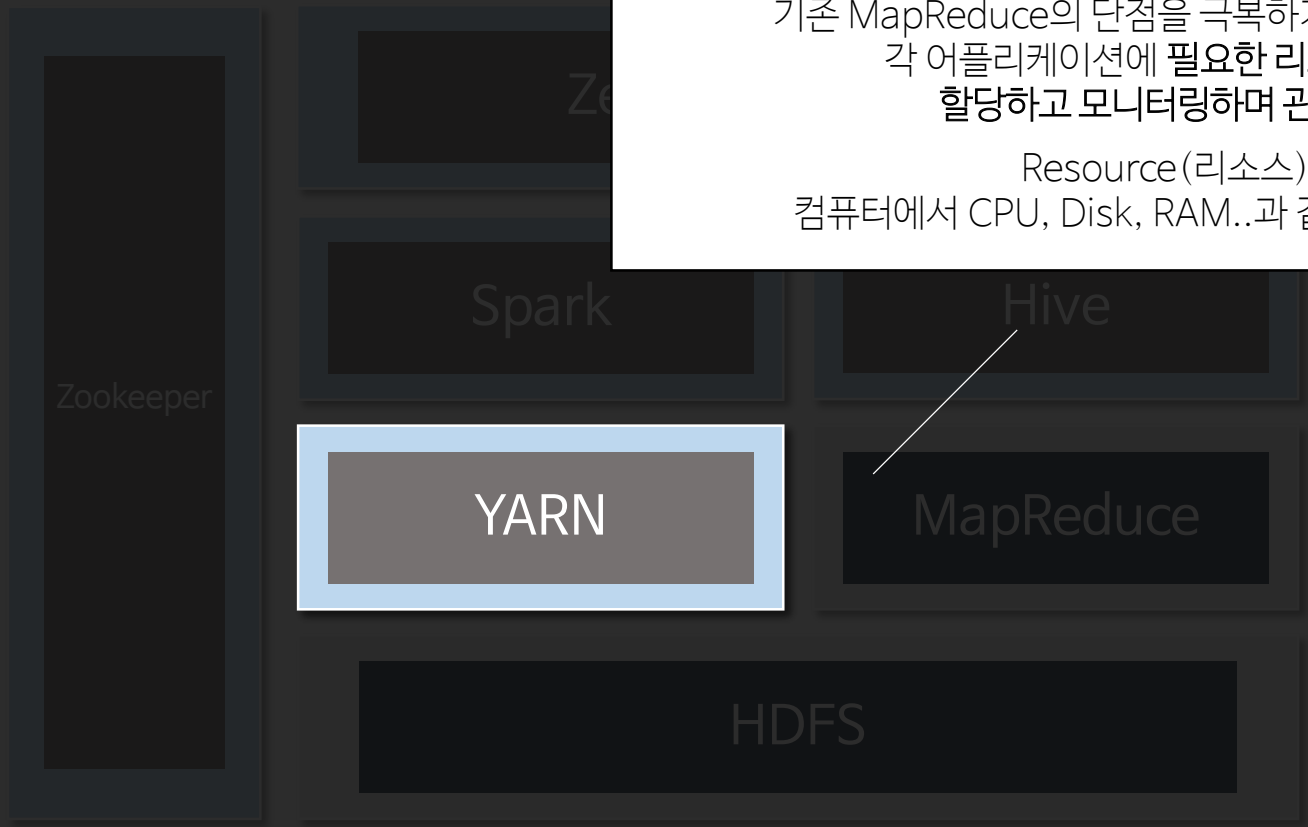
Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -

● **시스템의 리소스를 관리한다!**

기존 MapReduce의 단점을 극복하기 위해 만들어짐
각 어플리케이션에 **필요한 리소스들을**
할당하고 모니터링하며 관리함

Resource(리소스)?
컴퓨터에서 CPU, Disk, RAM..과 같은 자원을 뜻함!



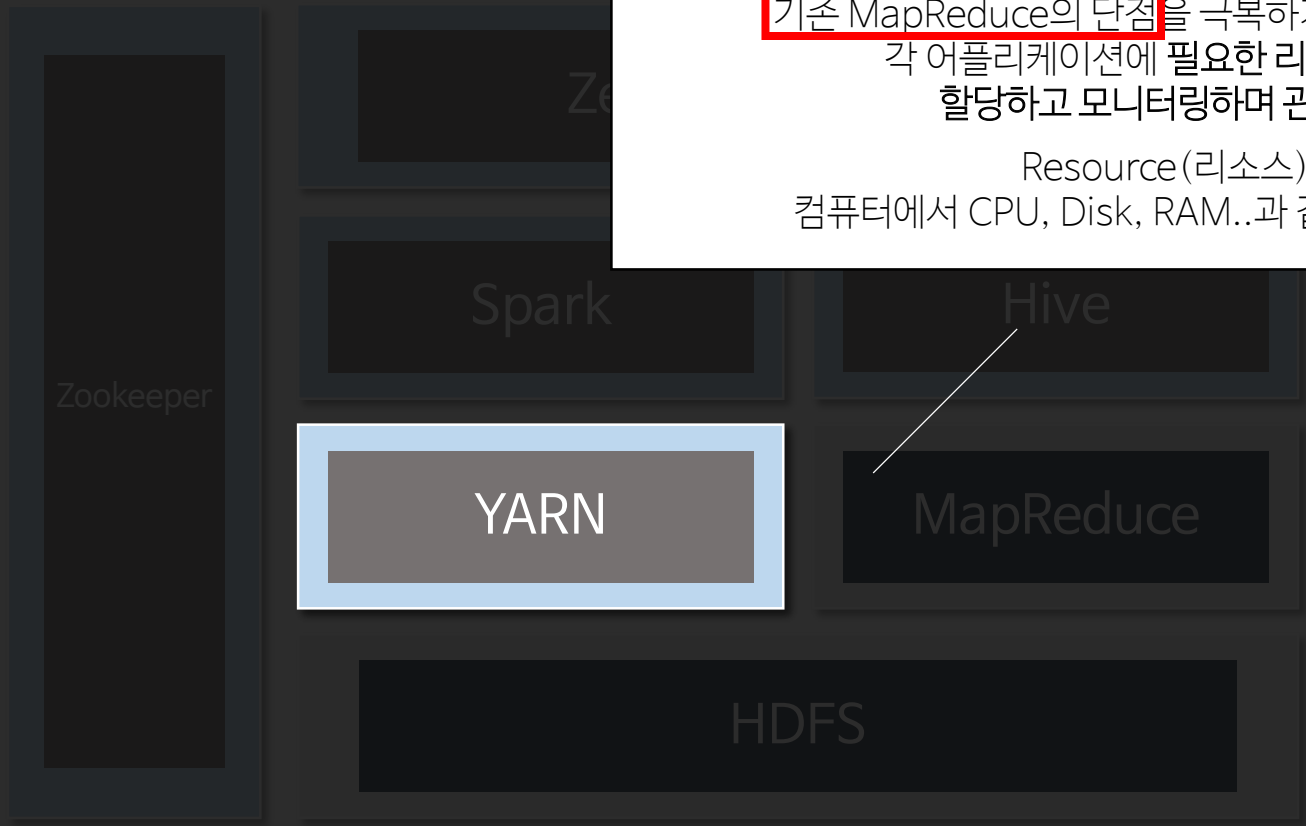
Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -

● **시스템의 리소스를 관리한다!**

기존 MapReduce의 단점을 극복하기 위해 만들어짐
각 어플리케이션에 **필요한 리소스들을**
할당하고 모니터링하며 관리함

Resource(리소스)?
컴퓨터에서 CPU, Disk, RAM..과 같은 자원을 뜻함!



MapRe

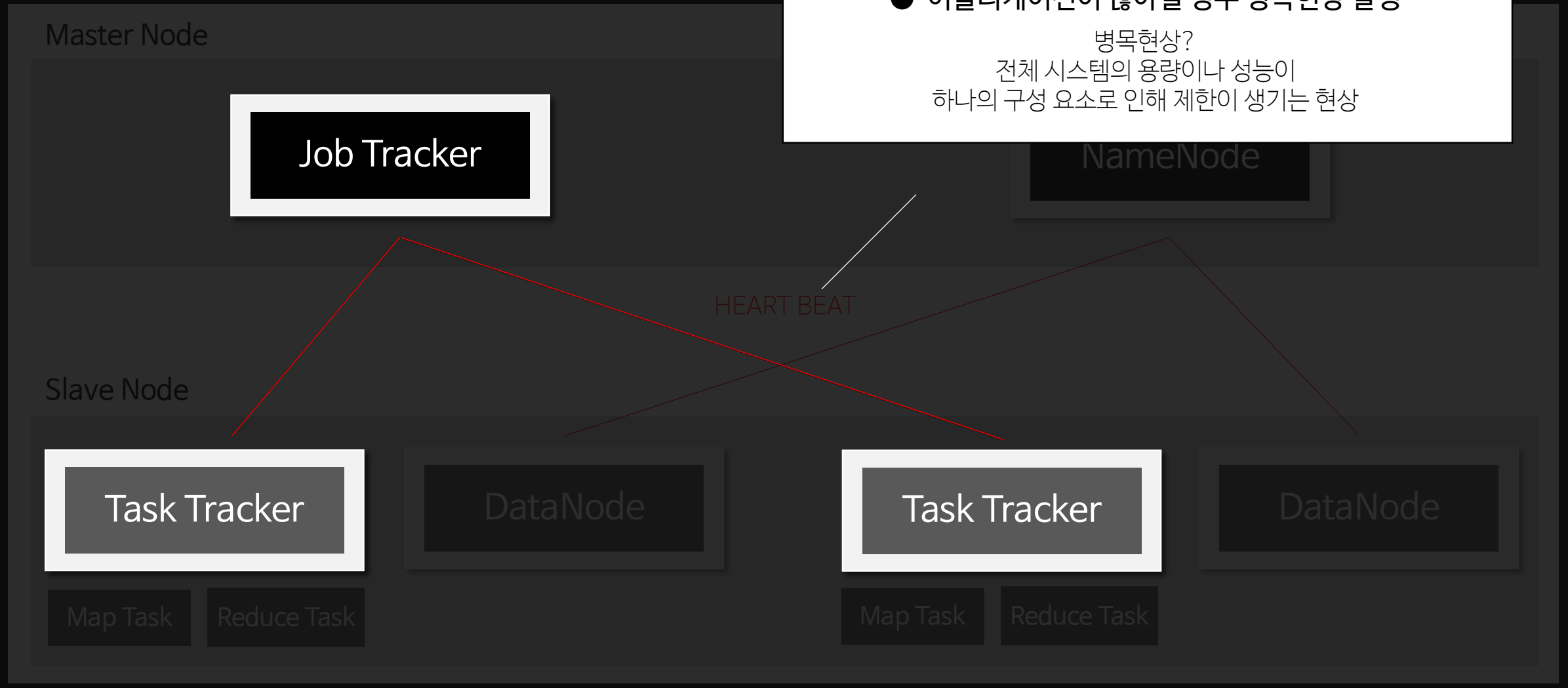
- 맵리듀스

● 하나의 Job Tracker가 모든 자원을 관리하는 문제

Job Tracker에 장애가 생기면 모든 어플리케이션에 장애 발생!

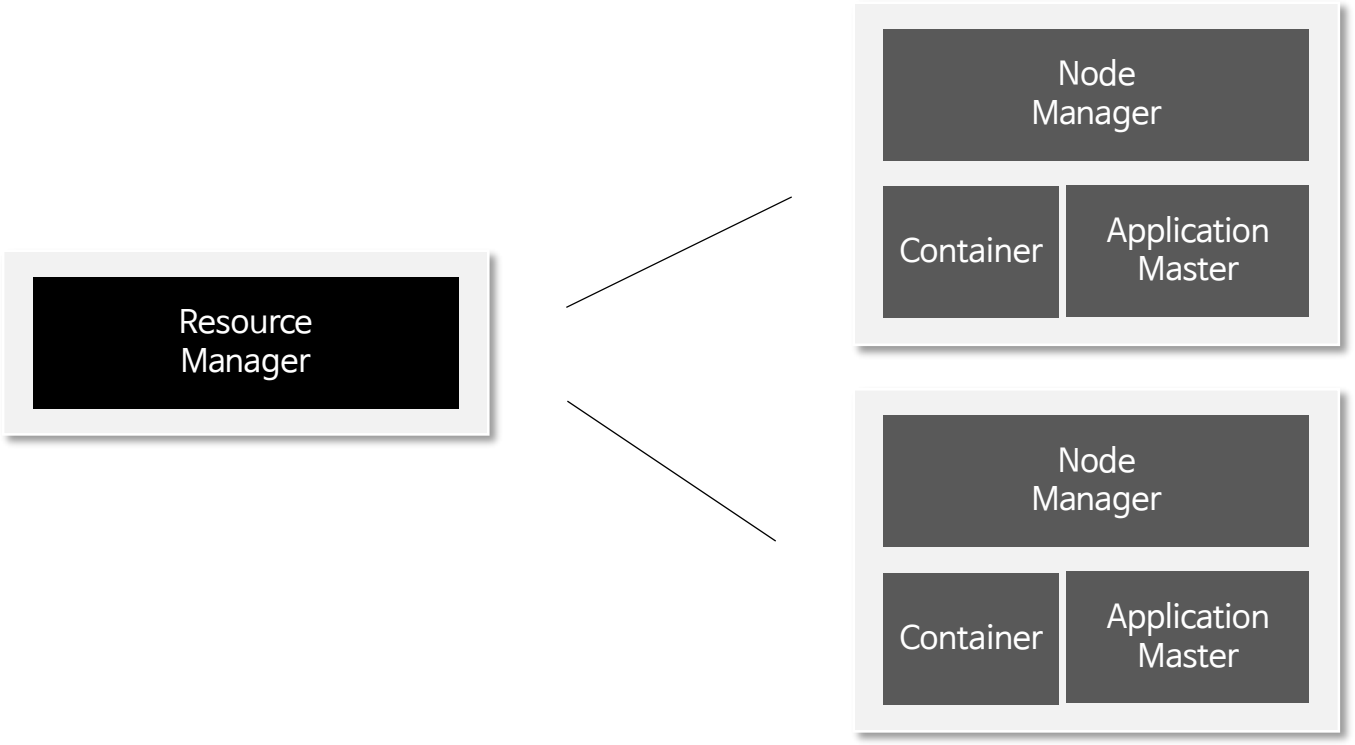
● 어플리케이션이 많아질 경우 병목현상 발생

병목현상?
전체 시스템의 용량이나 성능이
하나의 구성 요소로 인해 제한이 생기는 현상



YARN?

- YARN의 구성도 -



YARN?

- YARN의 구성도 -

Resource Manager?

YARN?

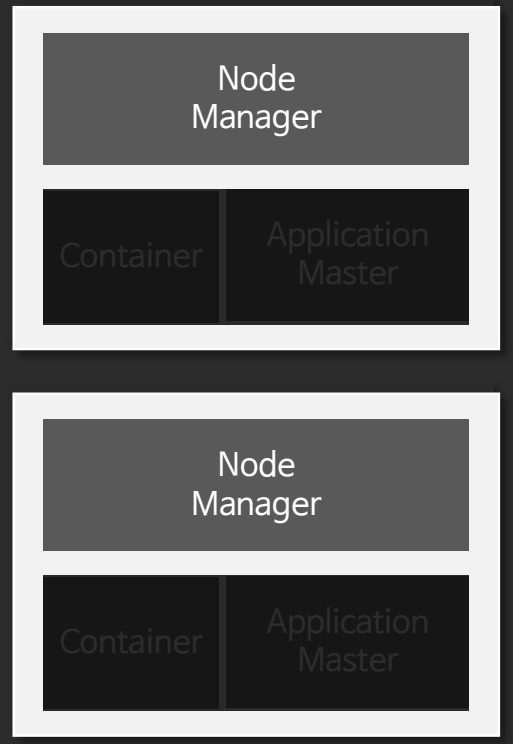
● Node Manager들을 컨트롤하는 역할을 함!

Master Node에 존재하며, Slave Node들에 있는 리소스들을 트래킹하고, Node Manager에게 Job을 할당하고 관리함

YARN?

- YARN의 구성도 -

Node Manager?

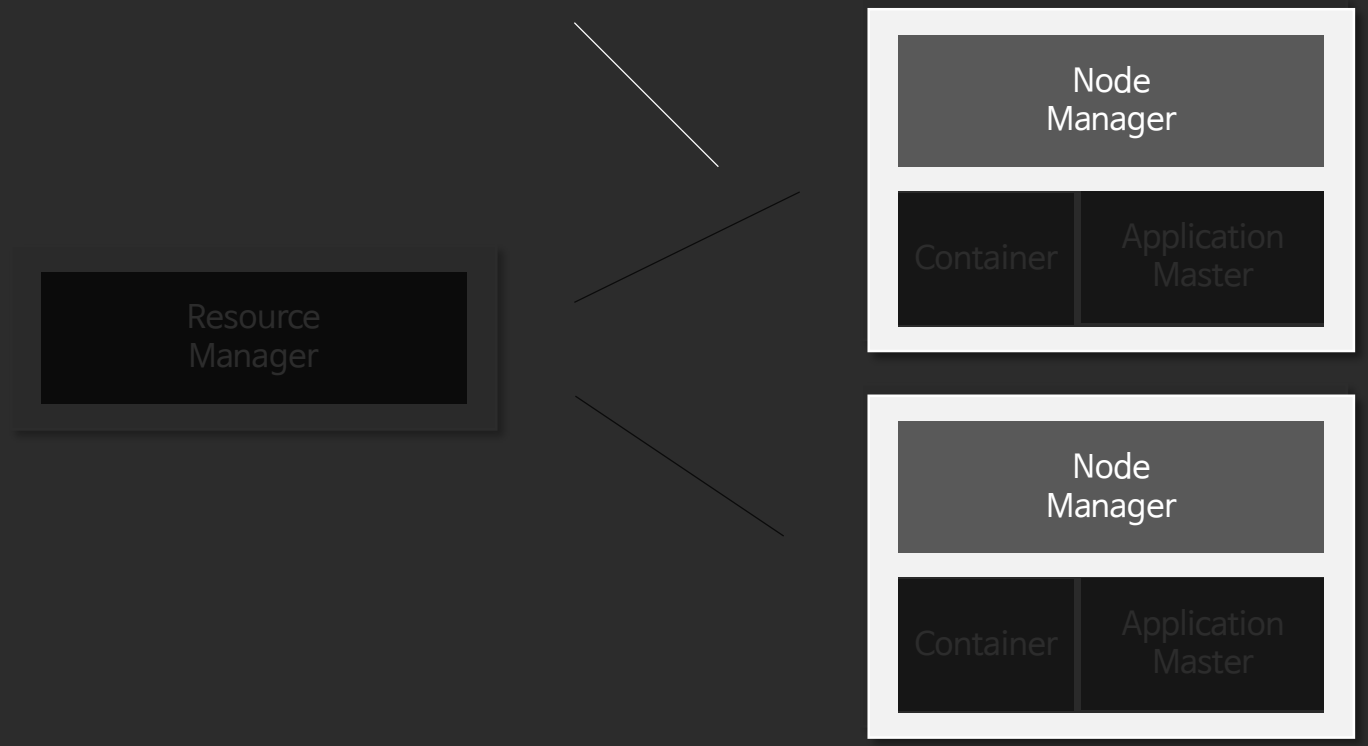


YARN?

- 구성도 -

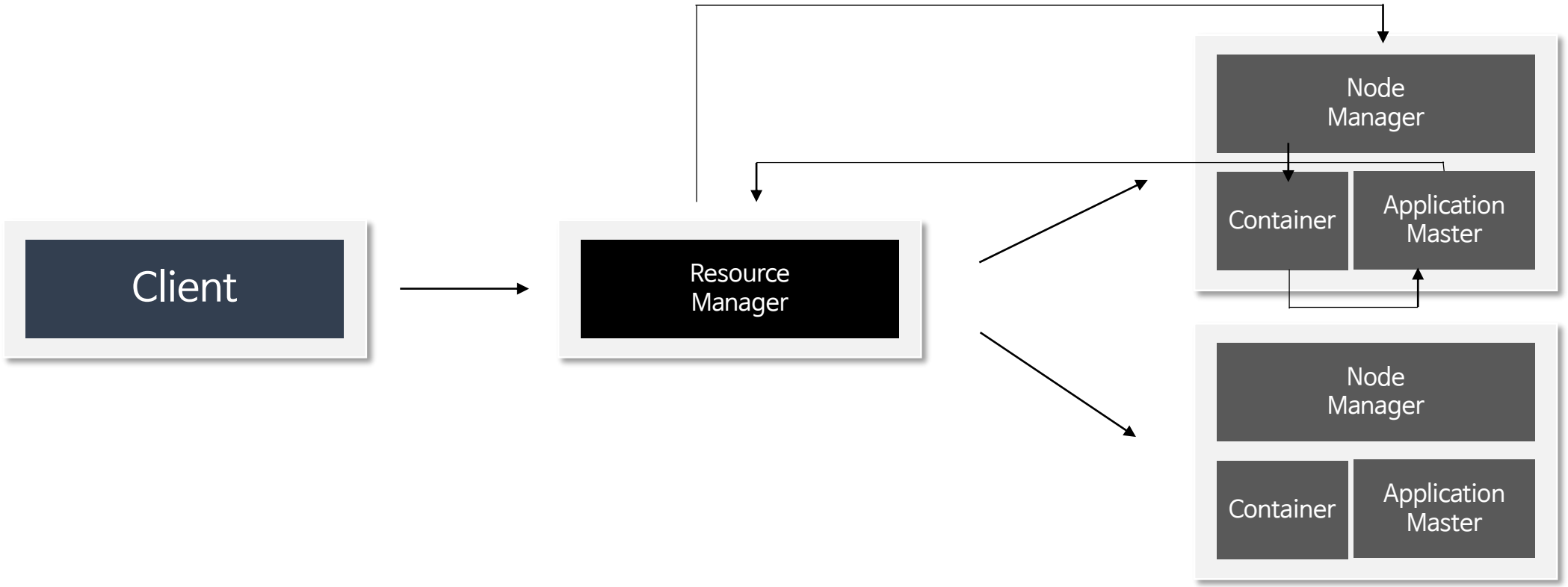
● Resource Manager에게 Job을 할당받고 수행함!

Slave Node에 존재하며 실제로 할당받은
Job을 처리하고, 결과를 전달함!



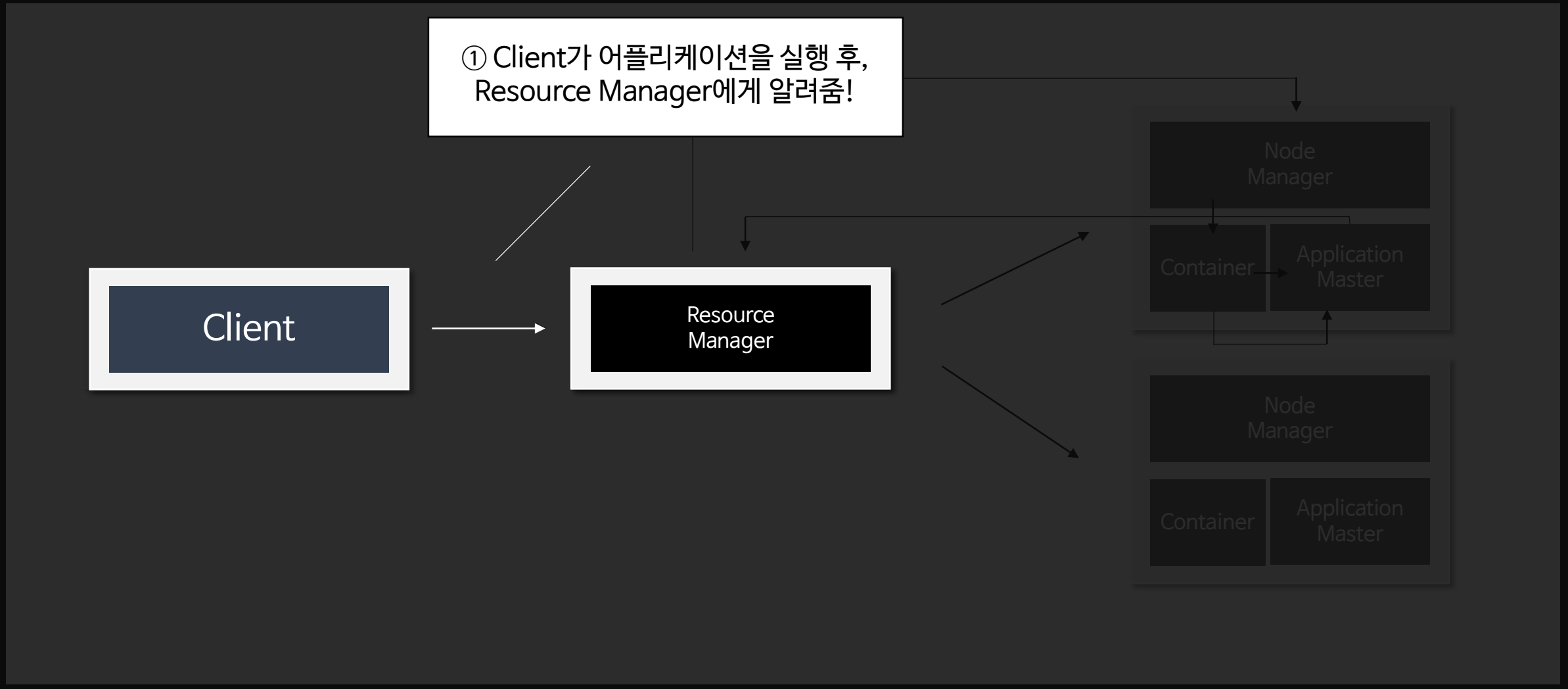
YARN?

- YARN의 Running Process -



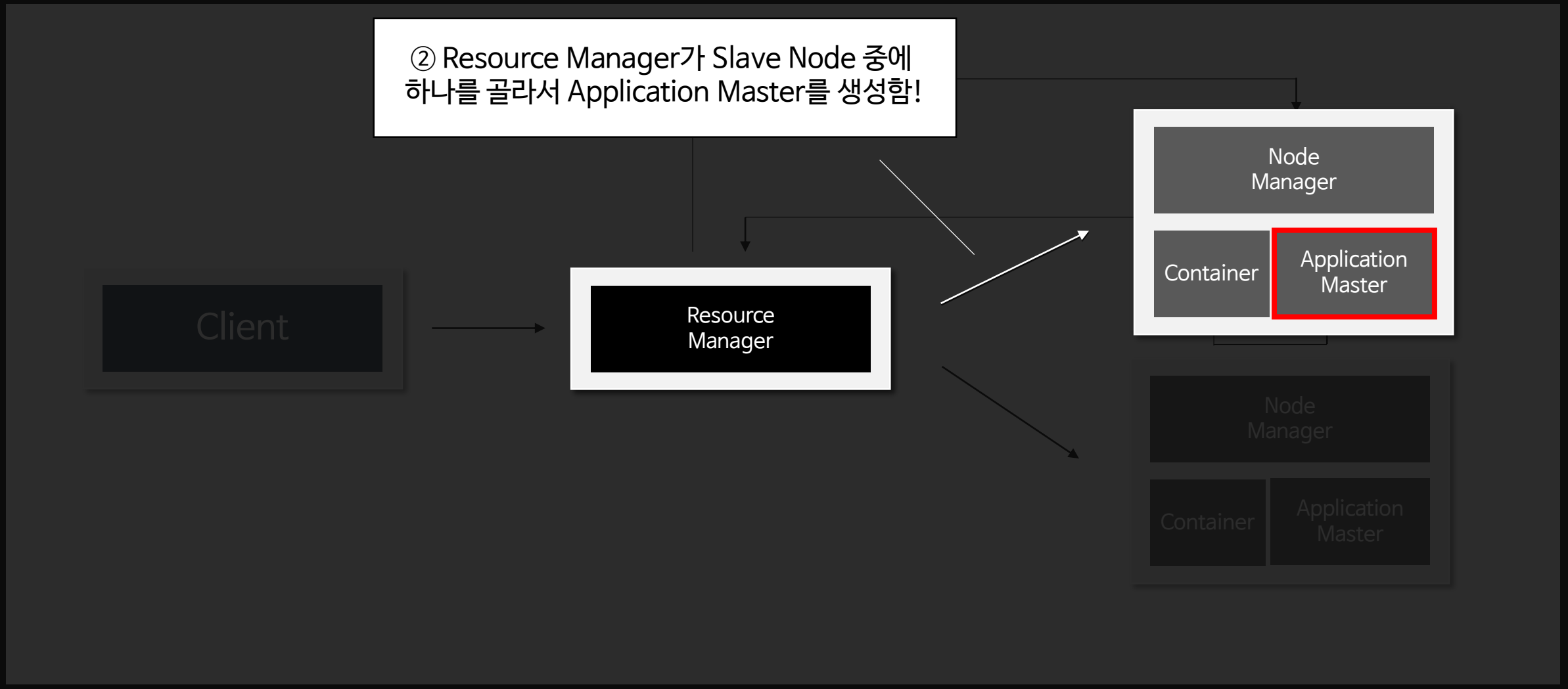
YARN?

- YARN의 Running Process -



YARN?

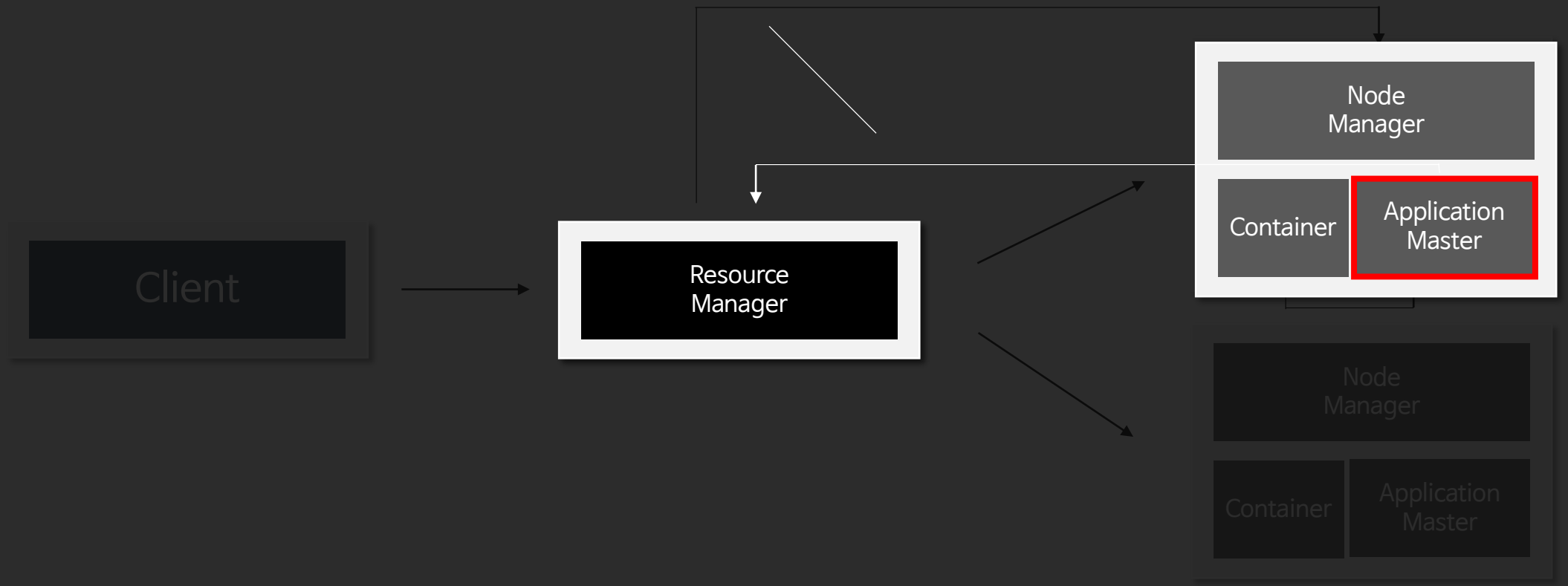
- YARN의 Running Process -



YARN?

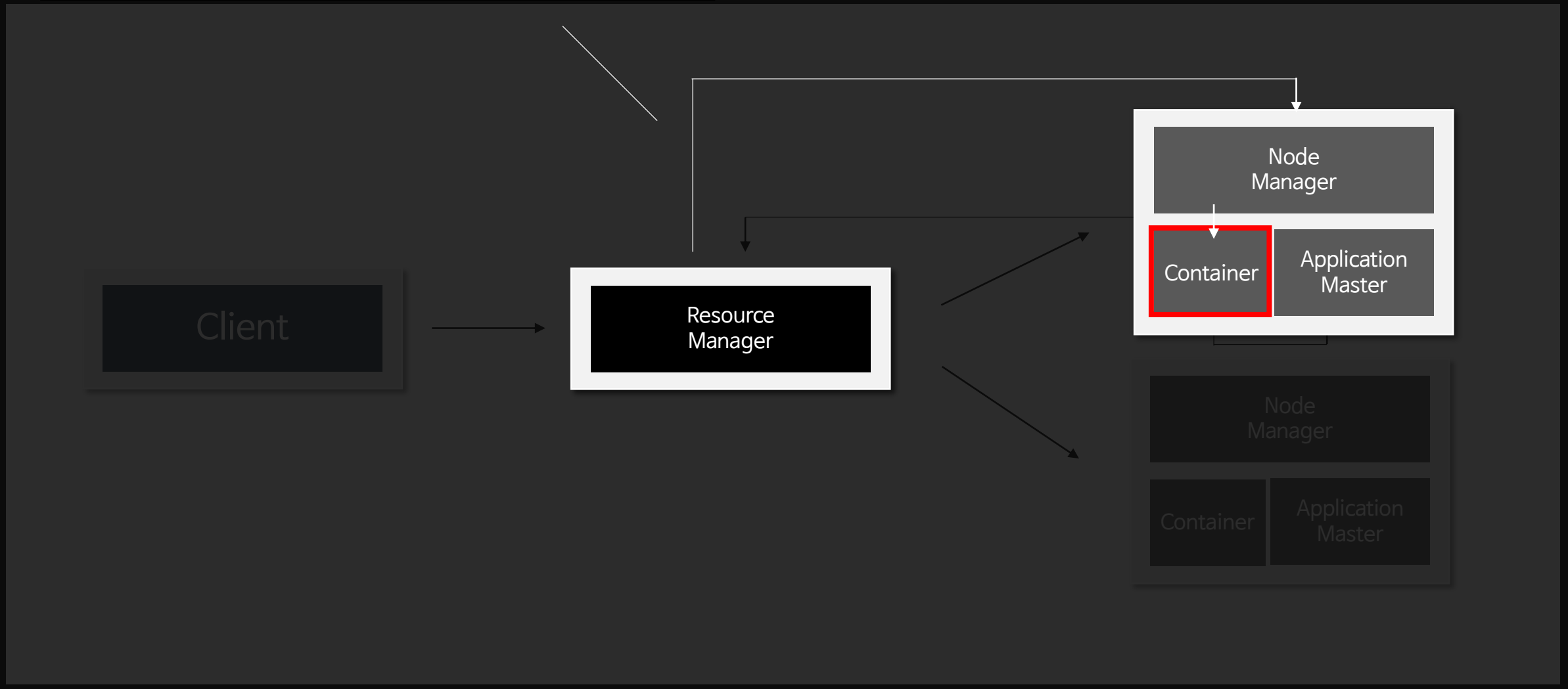
– YARN의 Running Process –

③ Application Master는 Resource Manager에게 Task를 수행할 컨테이너를 요청함!



④ Resource Manager는 넉넉한 자원을 소유한
Node Manager를 통해 Task를 실행할 Container 생성!

Job 할당!



YARN?

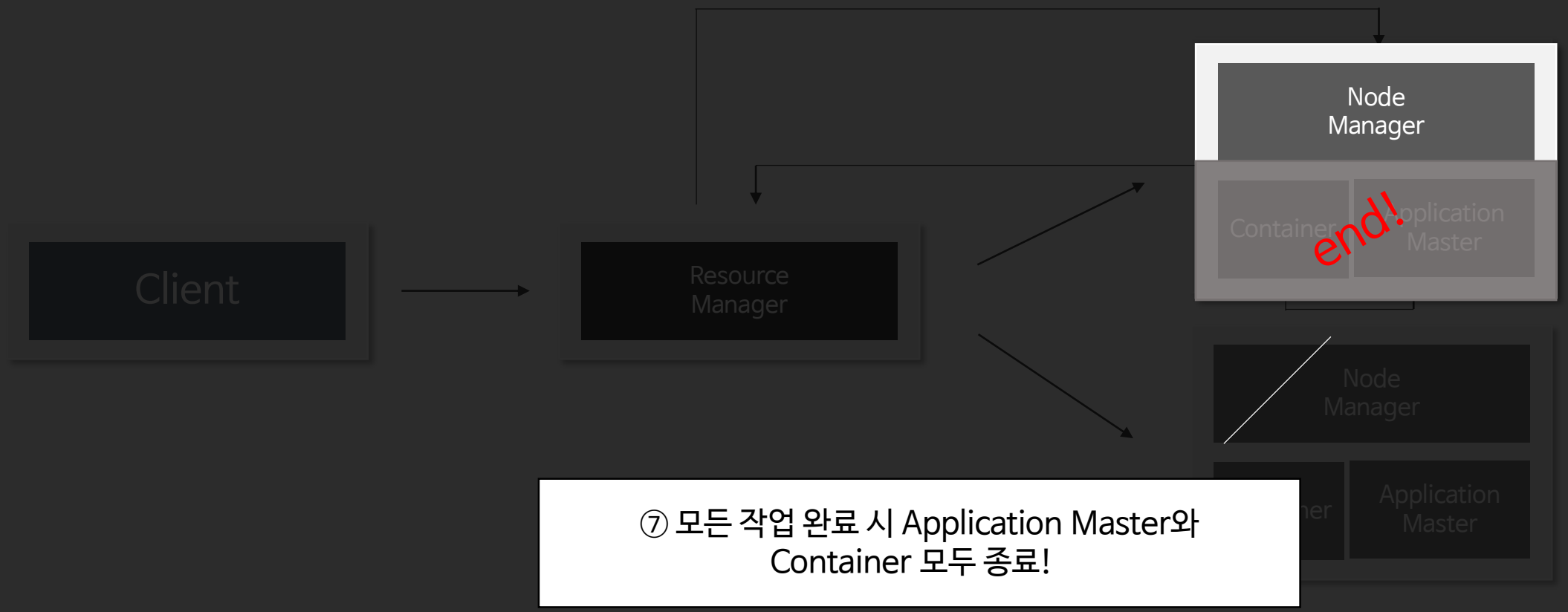
- YARN의 Running Process -

⑤ Task를 실행하는 Container는 Task의 상태를 Application Master에게 알려줌

Job 수행!

YARN?

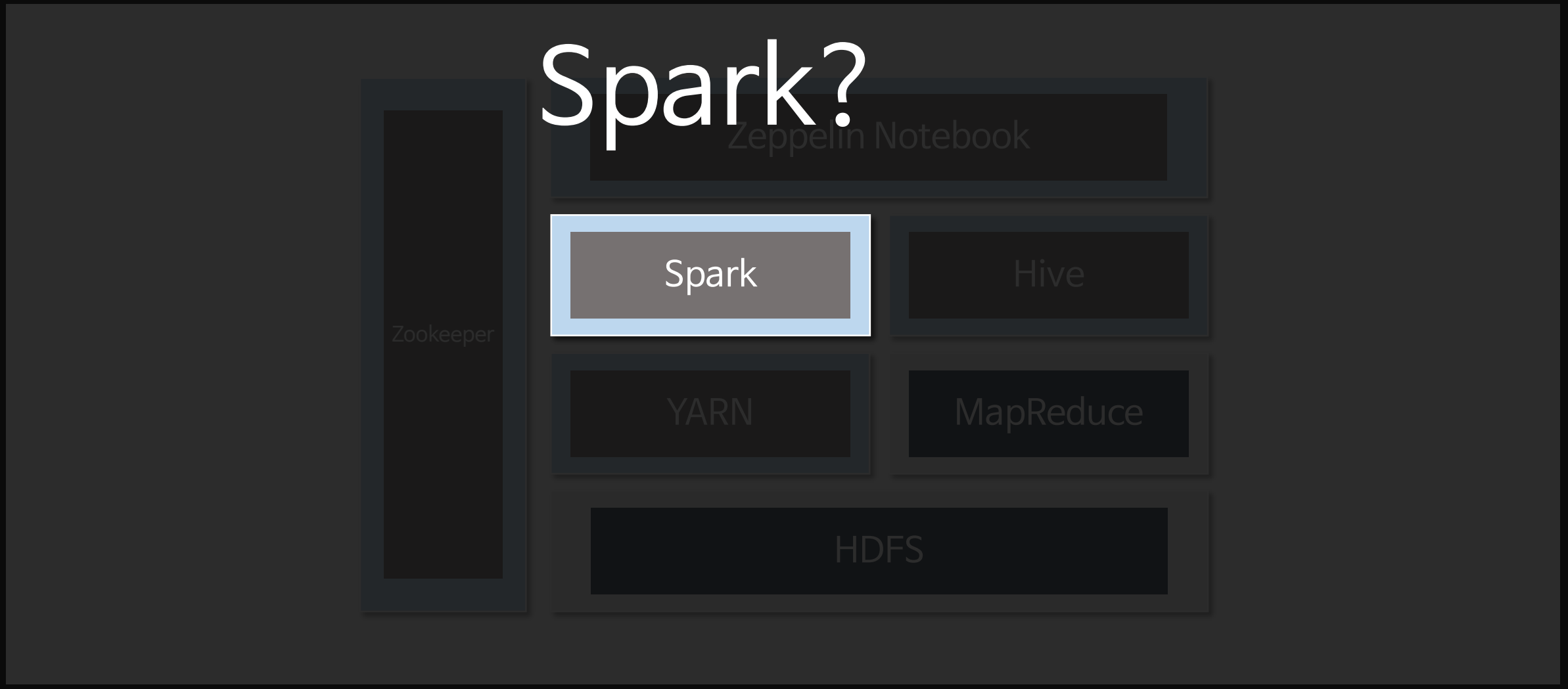
- YARN의 Running Process -



⑦ 모든 작업 완료 시 Application Master와 Container 모두 종료!

Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -



ecosystem?

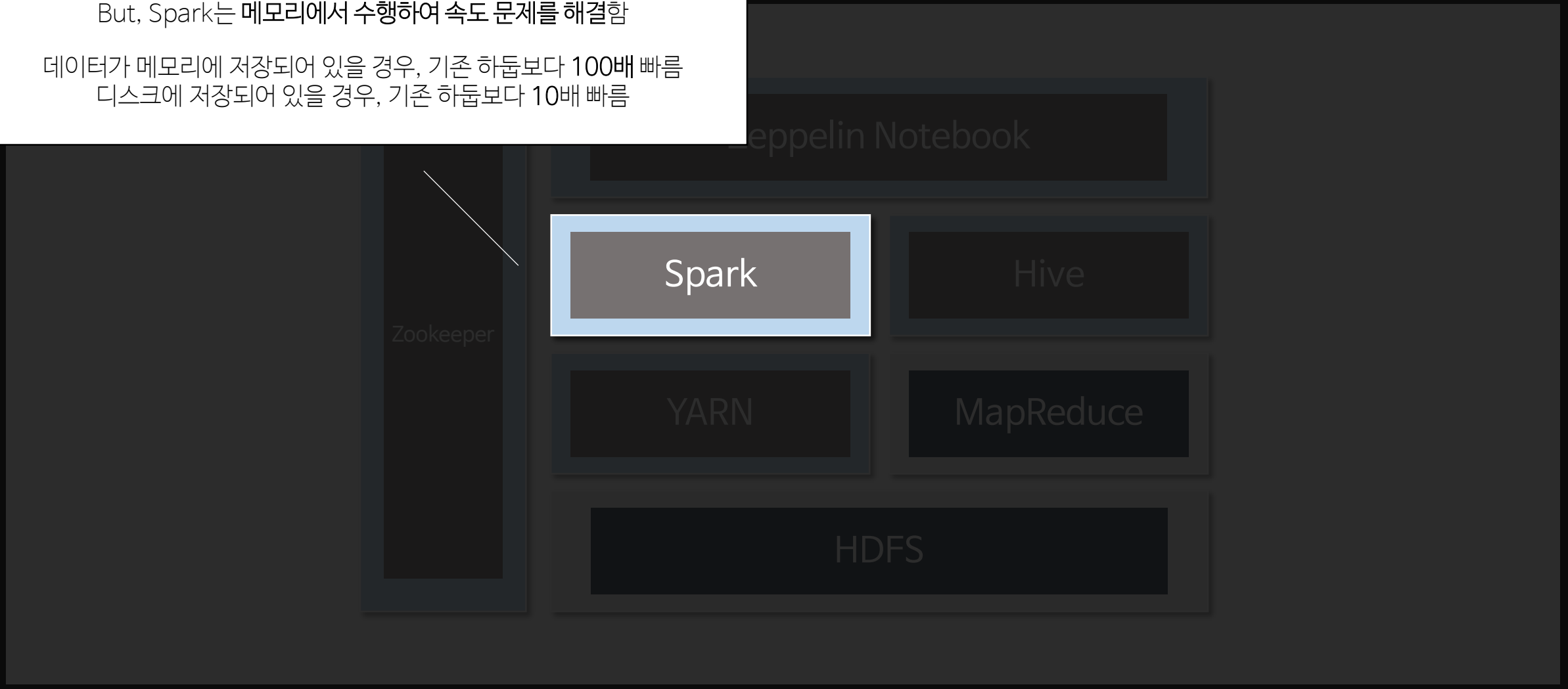
스템이란 무엇일까? –

● **범용적 목적 분산 고성능 처리 플랫폼!**

분산된 여러 대의 노드에서 빠르게 연산을 할 수 있게 해주는 플랫폼!

기존 하둡은 MapReduce를 디스크(HDFS)를 거쳐서 수행해서 속도가 느림
But, Spark는 **메모리에서 수행하여 속도 문제를 해결**함

데이터가 메모리에 저장되어 있을 경우, 기존 하둡보다 **100배** 빠름
디스크에 저장되어 있을 경우, 기존 하둡보다 10배 빠름



ecosystem?

스템이란 무엇일까? –

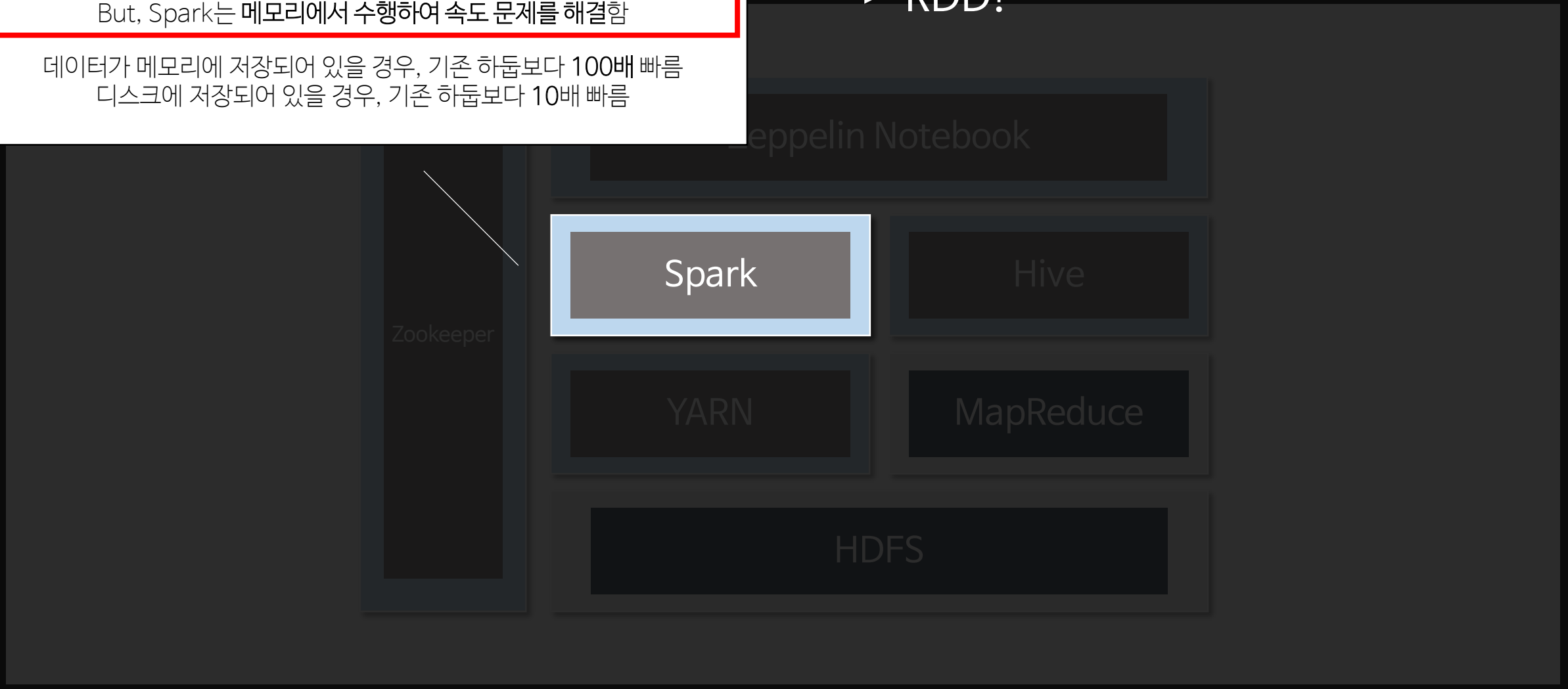
● **범용적 목적 분산 고성능 처리 플랫폼!**

분산된 여러 대의 노드에서 빠르게 연산을 할 수 있게 해주는 플랫폼!

기존 하둡은 MapReduce를 디스크(HDFS)를 거쳐서 수행해서 속도가 느림
But, Spark는 **메모리에서 수행하여 속도 문제를 해결**함

데이터가 메모리에 저장되어 있을 경우, 기존 하둡보다 **100배** 빠름
디스크에 저장되어 있을 경우, 기존 하둡보다 10배 빠름

→ RDD!



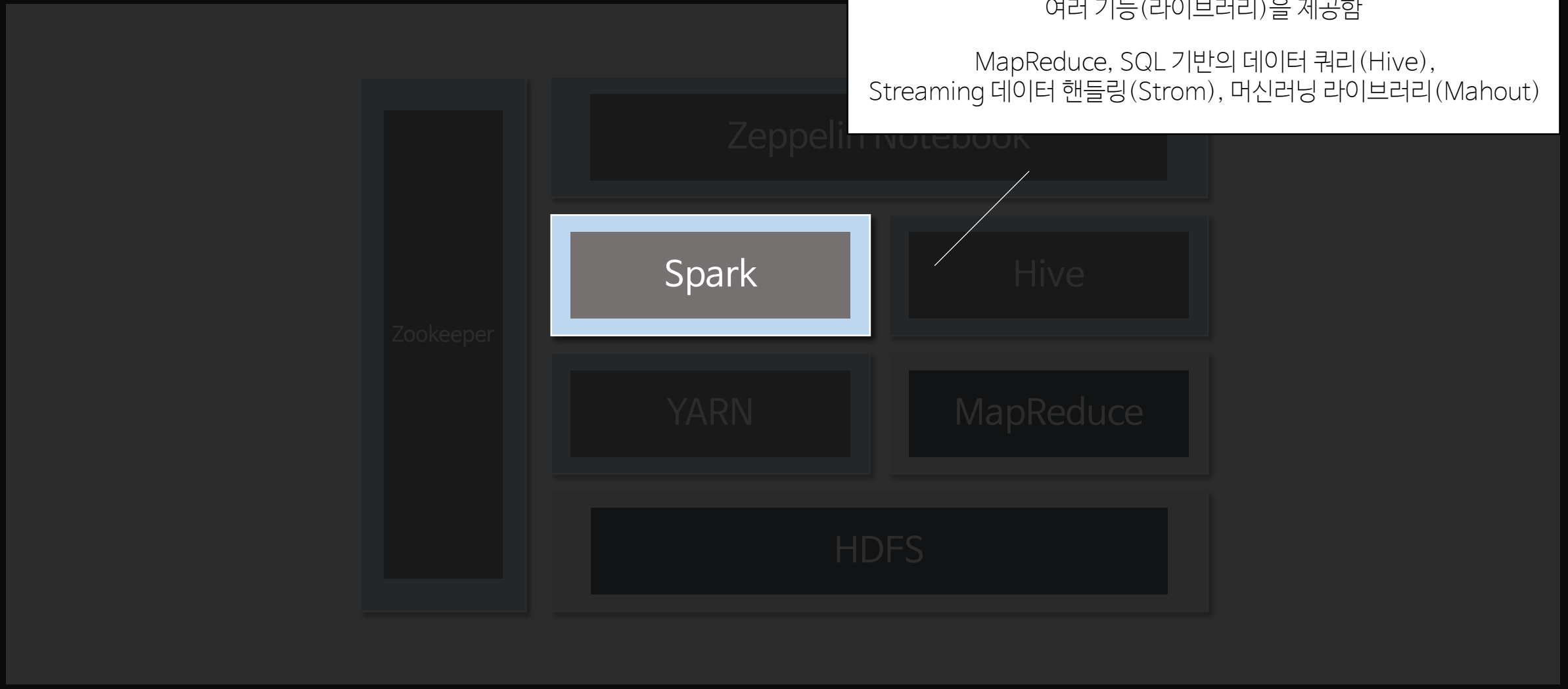
Hadoop ecosystem?

– 하둡 에코시스템이란

● **“범용적 목적”이므로 여러 기능 제공!**

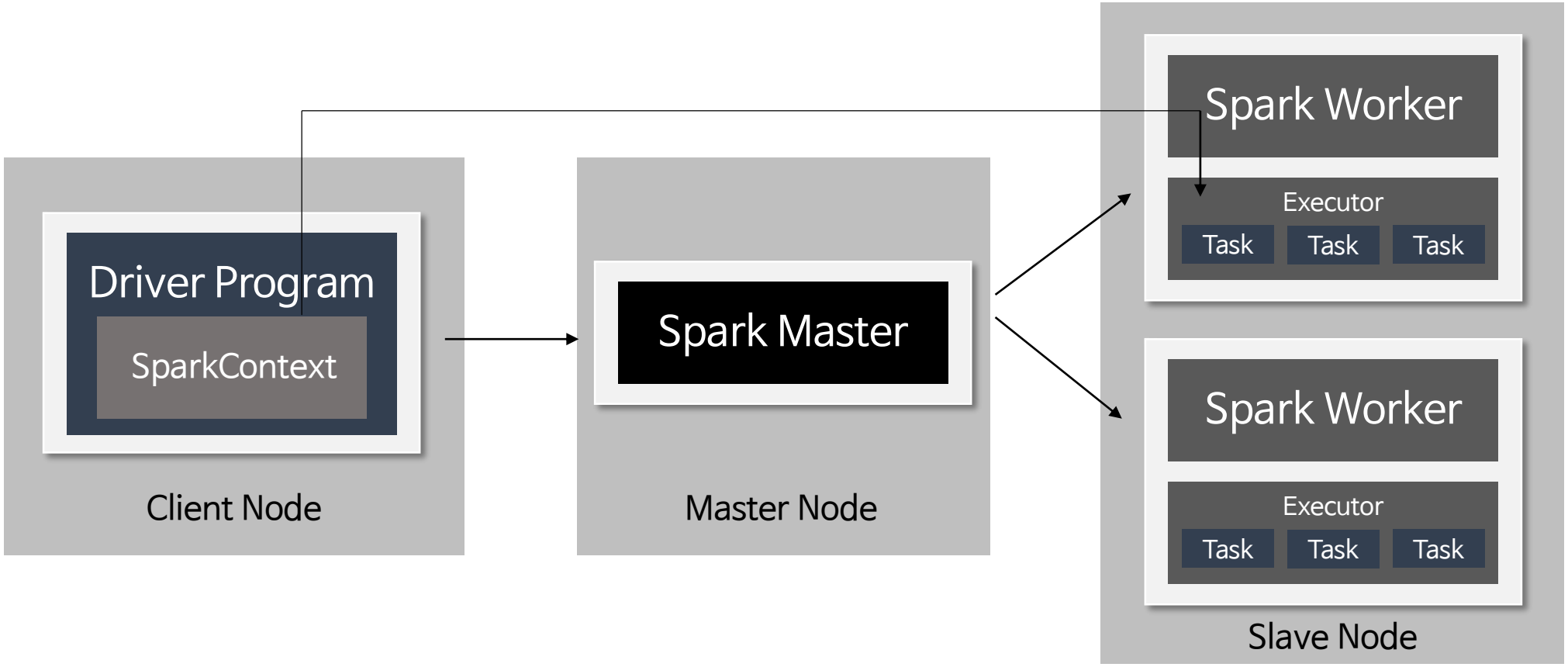
기존 하둡처럼 MapReduce 모듈만 돌리지 않고,
여러 기능(라이브러리)을 제공함

MapReduce, SQL 기반의 데이터 쿼리(Hive),
Streaming 데이터 핸들링(Strom), 머신러닝 라이브러리(Mahout)



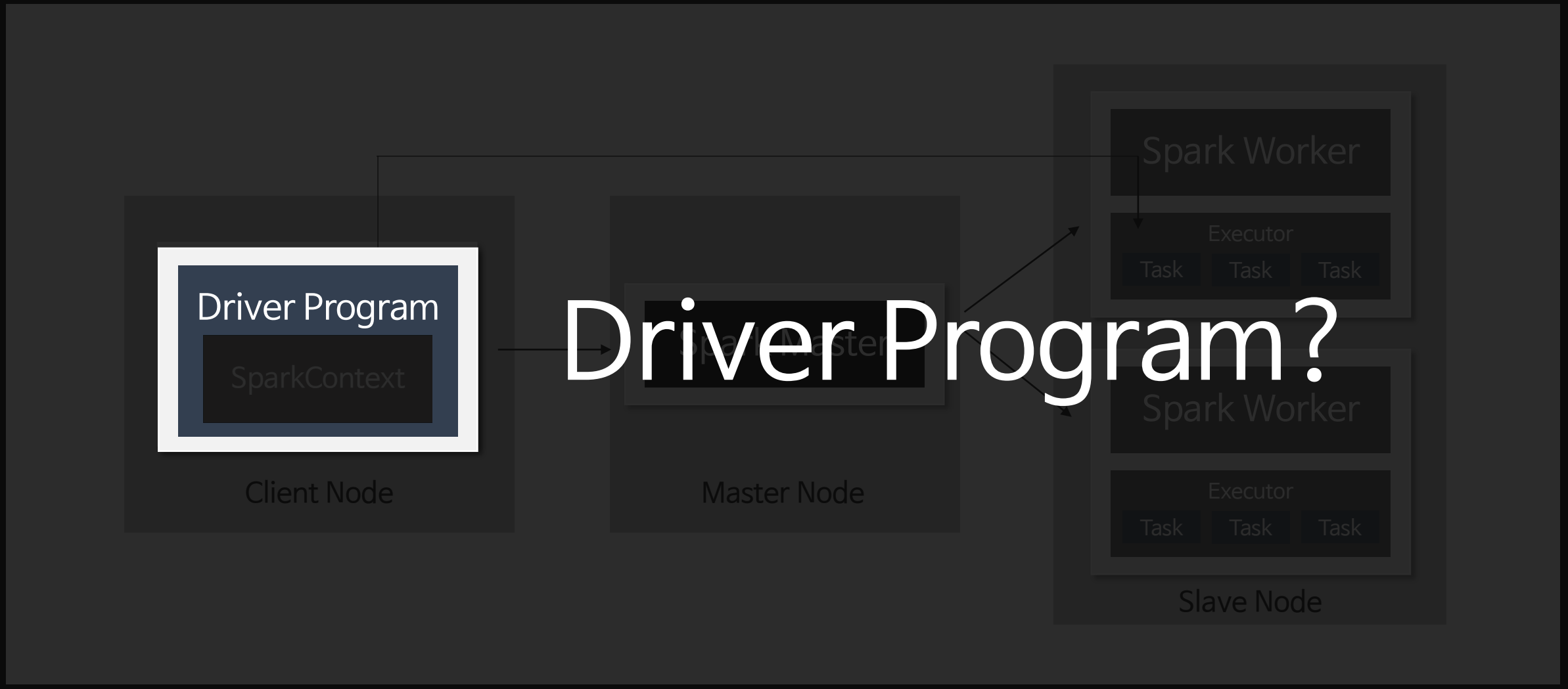
Spark?

- Spark 구성도 -



Spark?

- Spark 구성도 -



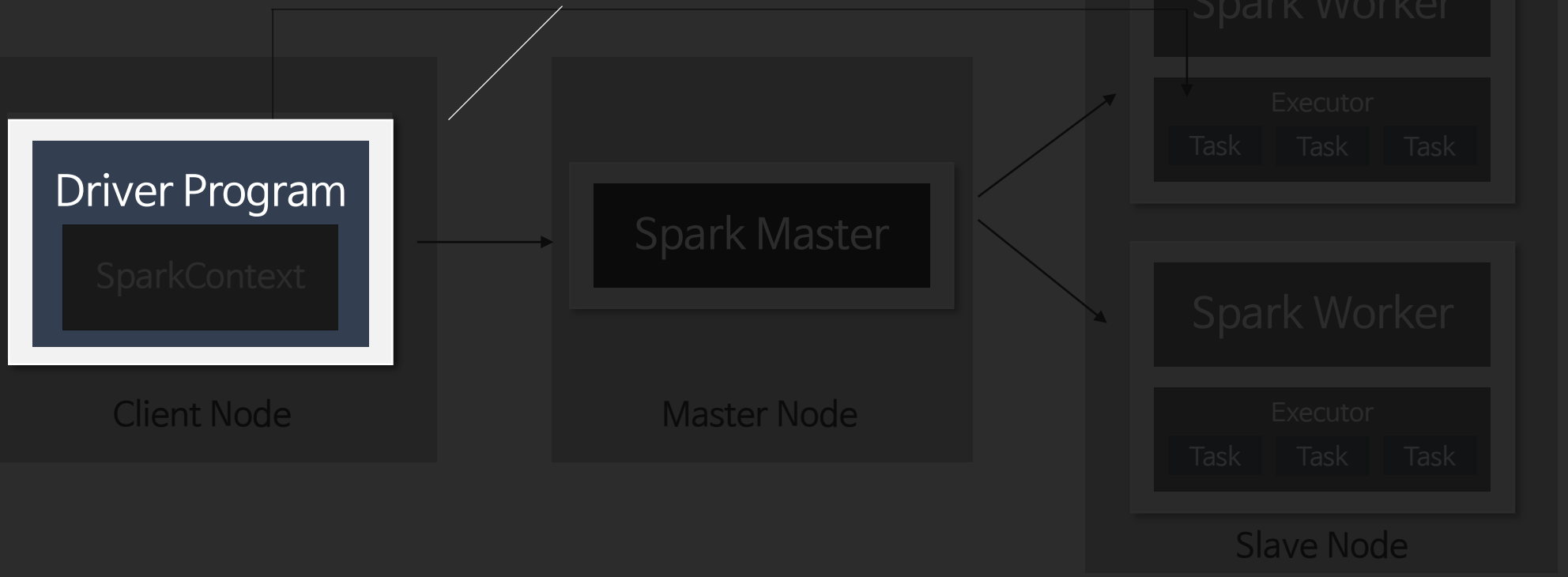
Driver Program?

Spark?

- Spark 구성도 -

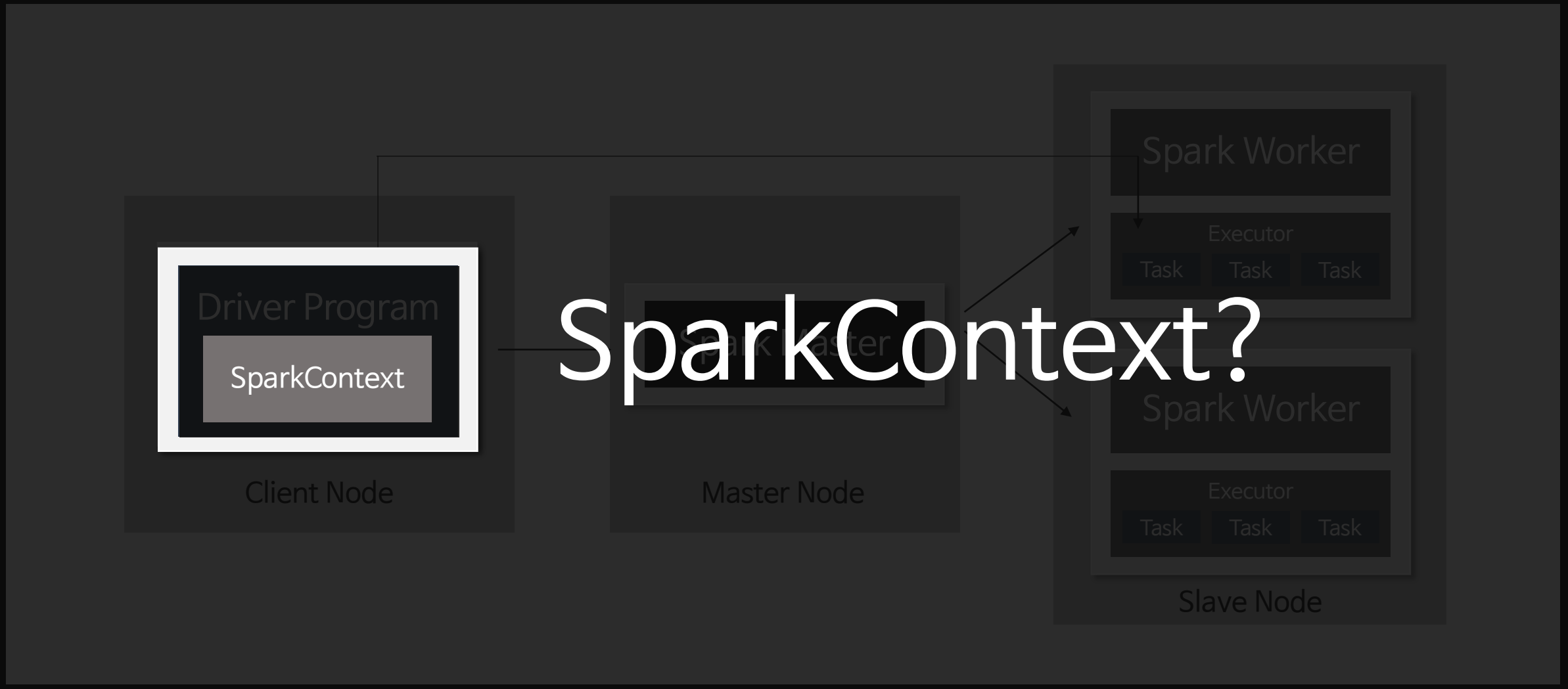
● 메인 프로세스

Main 함수가 실행되는 곳!
그 안에서 SparkContext를 생성함



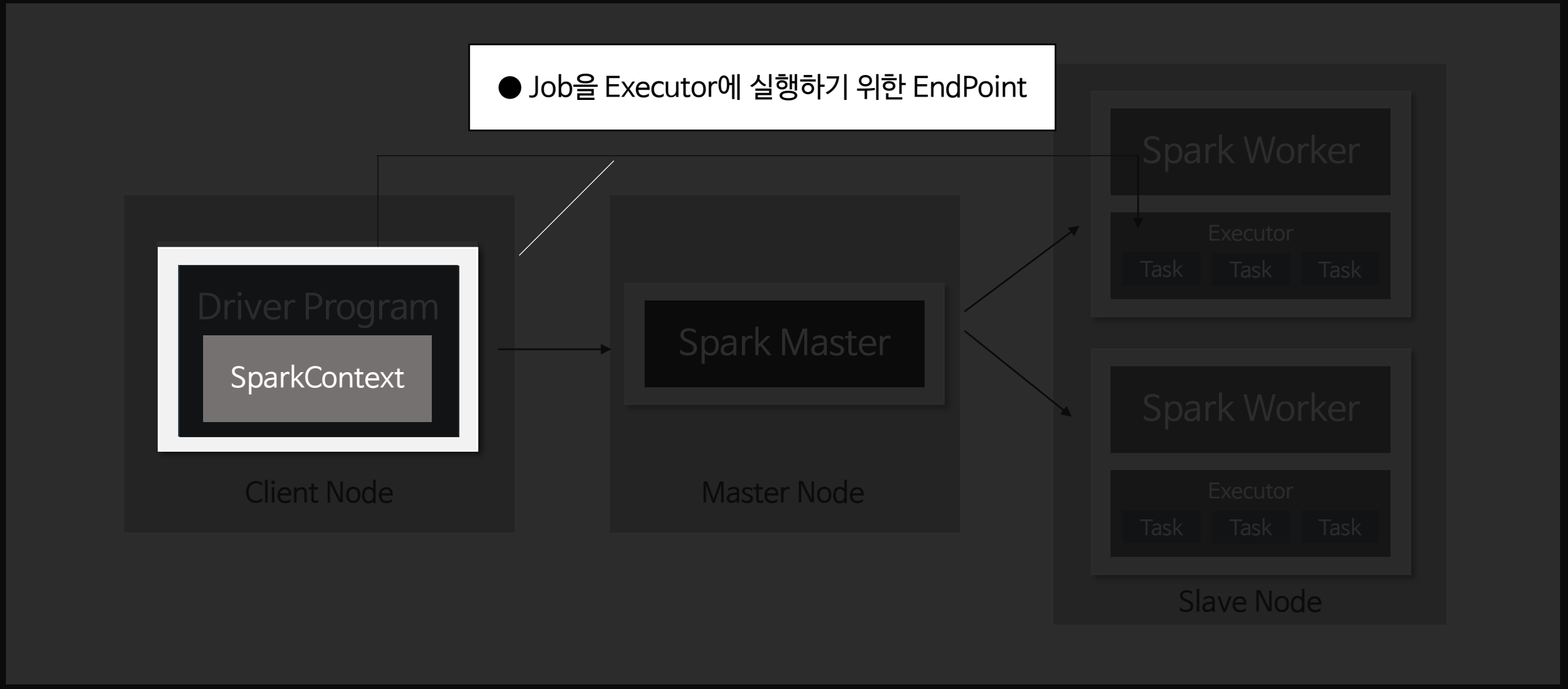
Spark?

- Spark 구성도 -



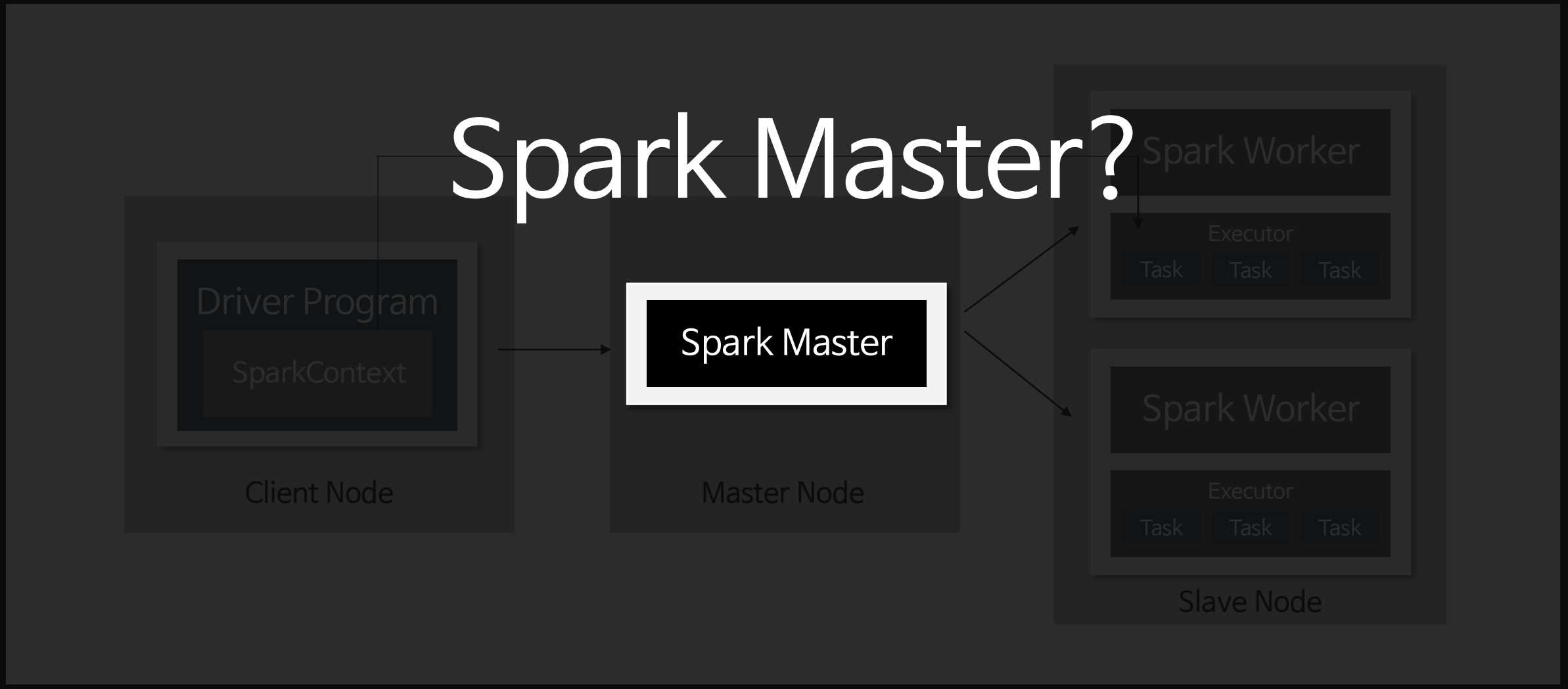
Spark?

- Spark 구성도 -



Spark?

- Spark 구성도 -

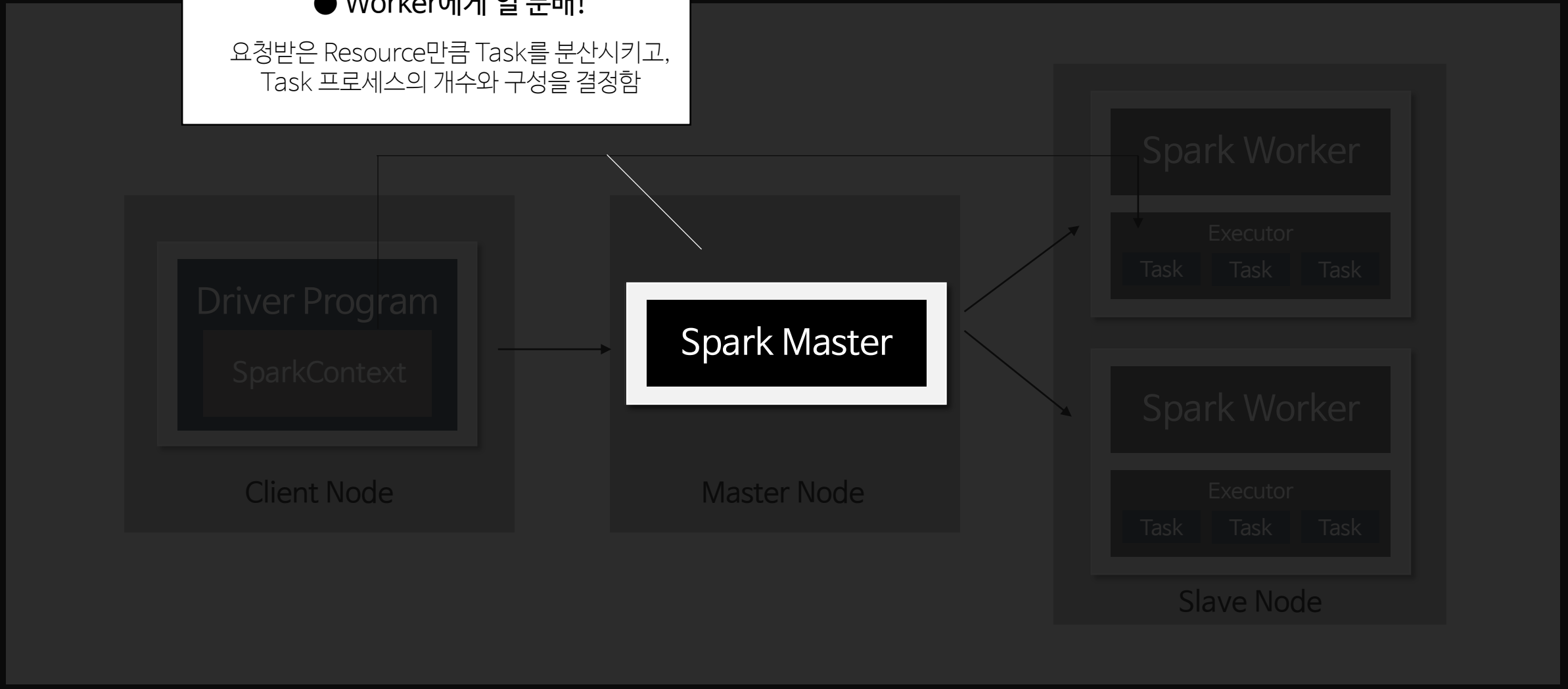


Spark?

- Spark 구성도 -

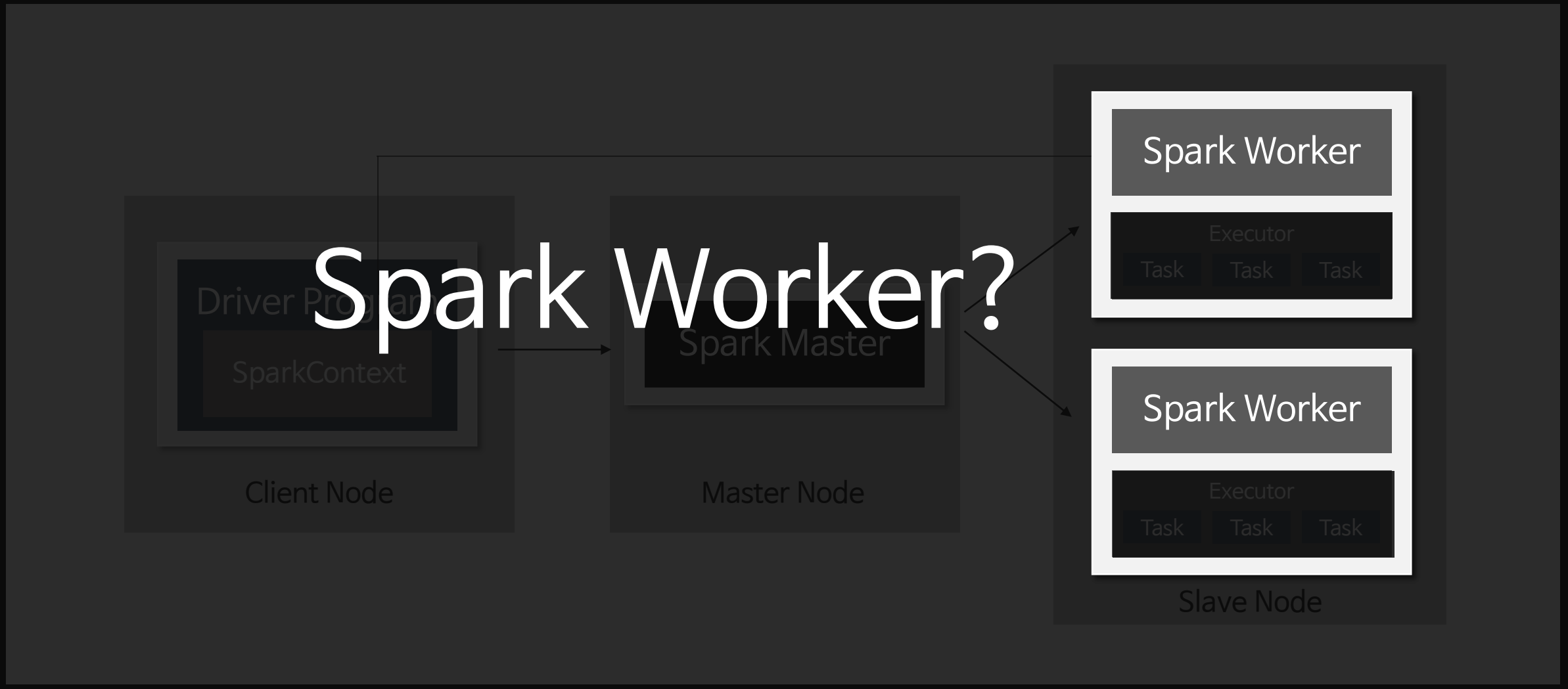
● Worker에게 일 분배!

요청받은 Resource만큼 Task를 분산시키고,
Task 프로세스의 개수와 구성을 결정함



Spark?

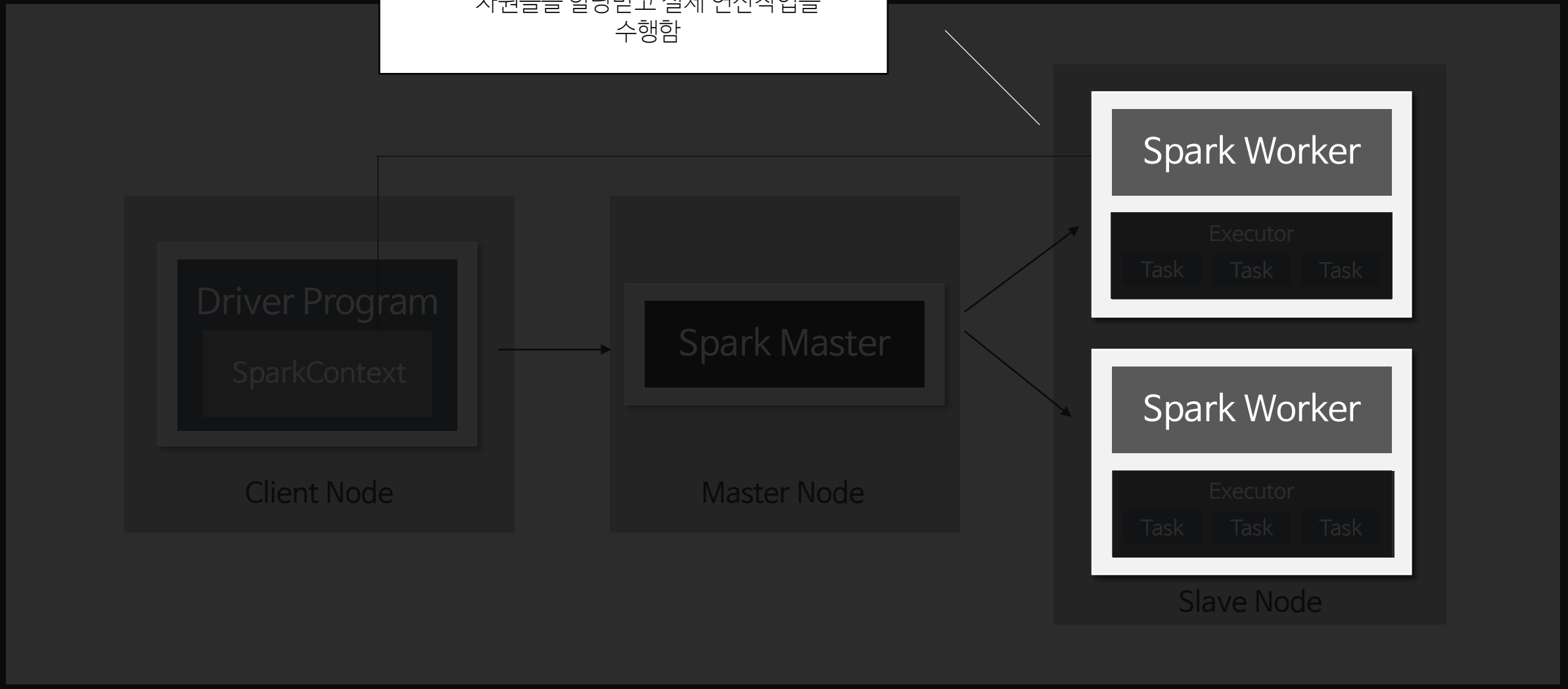
- Spark 구성도 -



Spark?

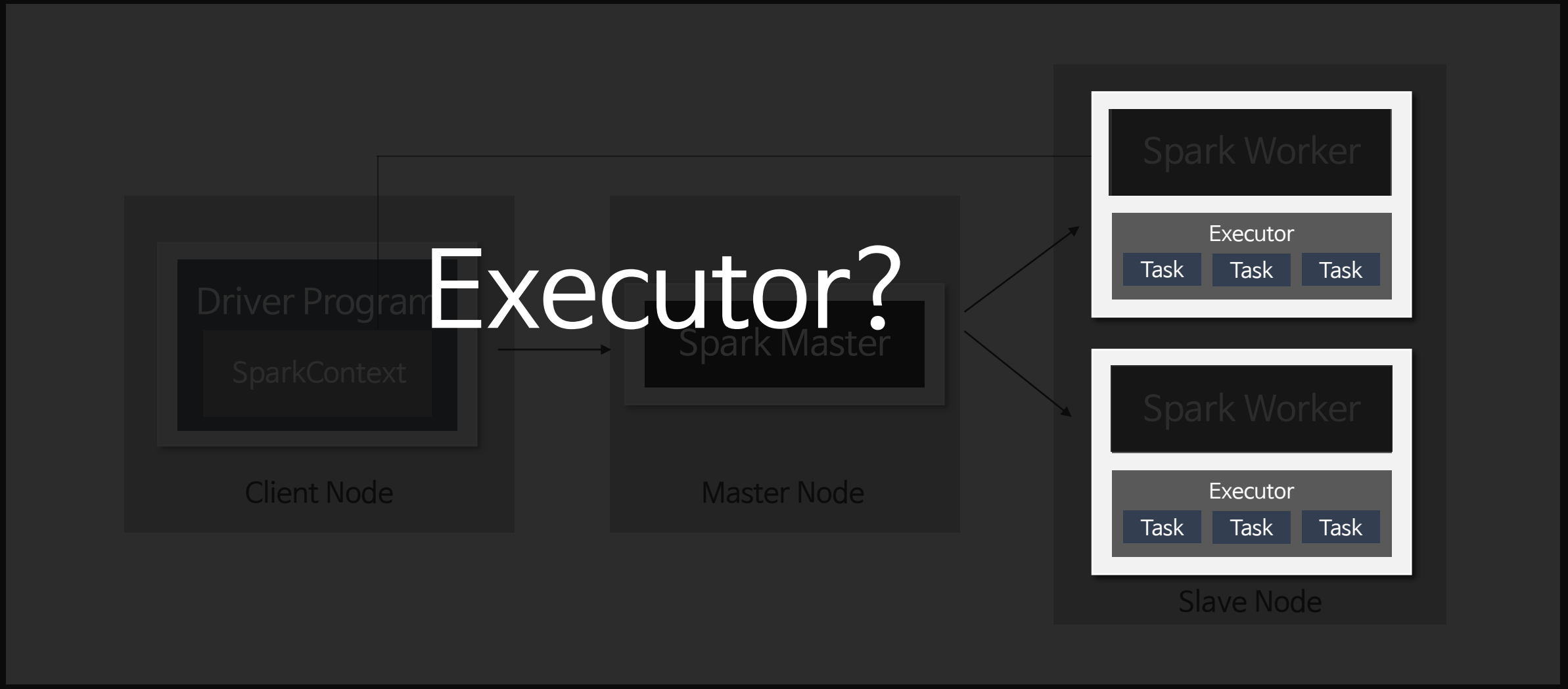
● 실제 연산작업을 수행하는 노드!

자원들을 할당받고 실제 연산작업을
수행함



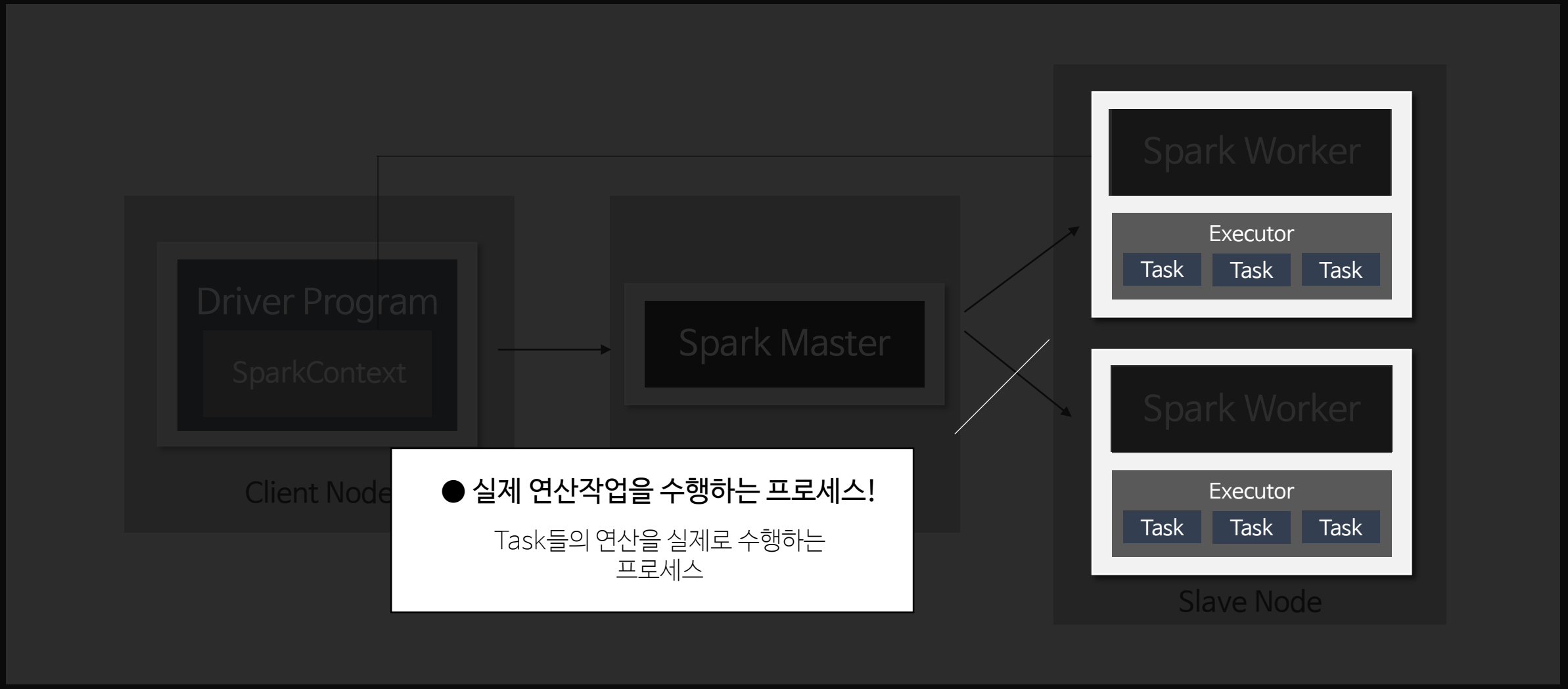
Spark?

- Spark 구성도 -



Spark?

- Spark 구성도 -



● 실제 연산작업을 수행하는 프로세스!

Task들의 연산을 실제로 수행하는
프로세스

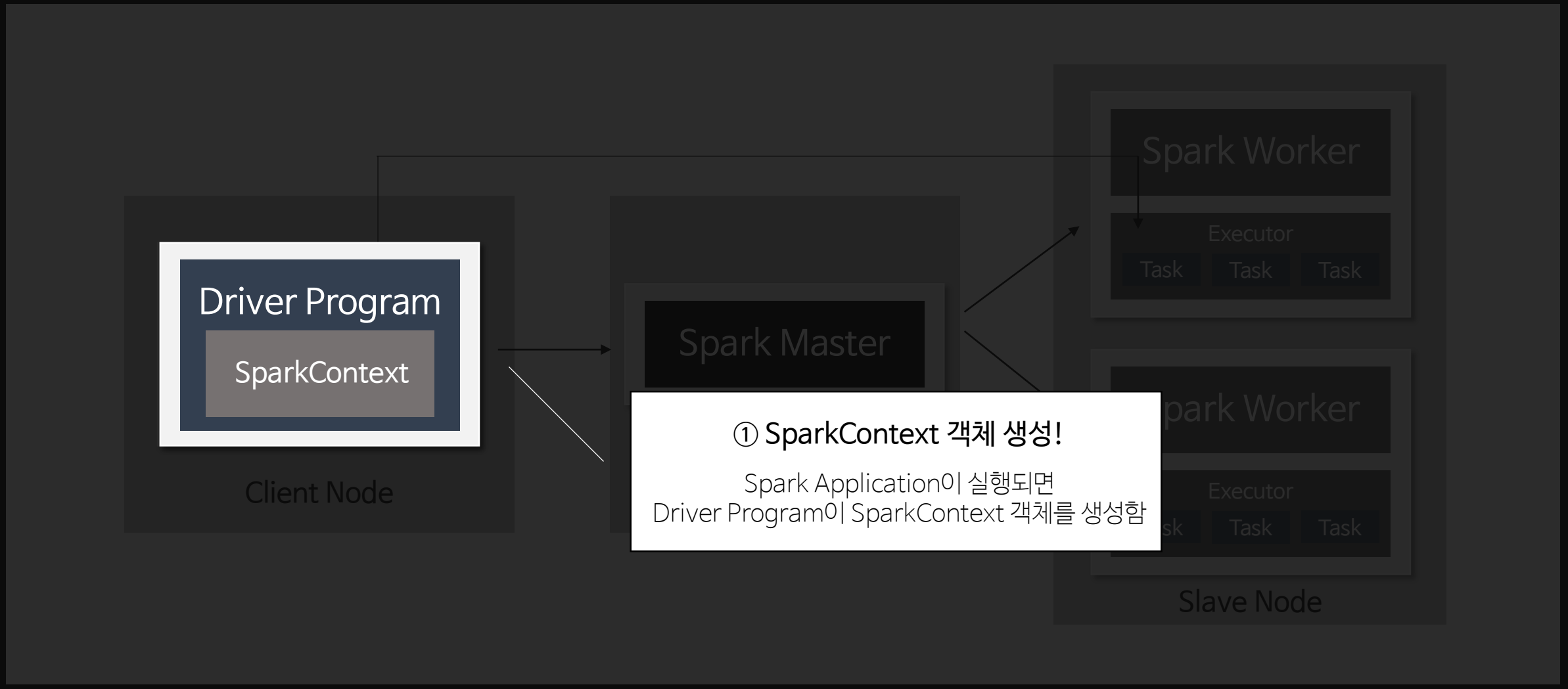
Spark?

- Spark 구성도 -

Running Process?

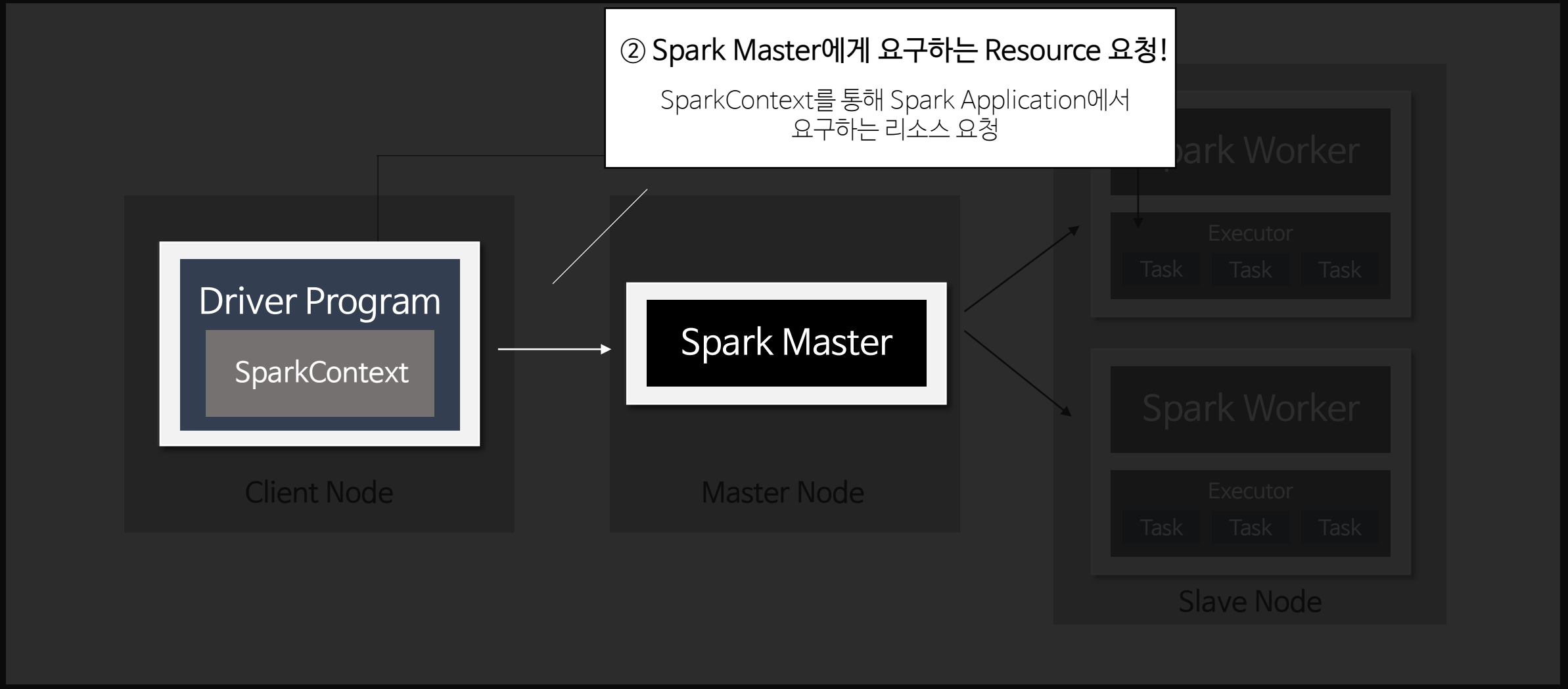
Spark?

- Spark 구성도 -



Spark?

- Spark 구성도 -

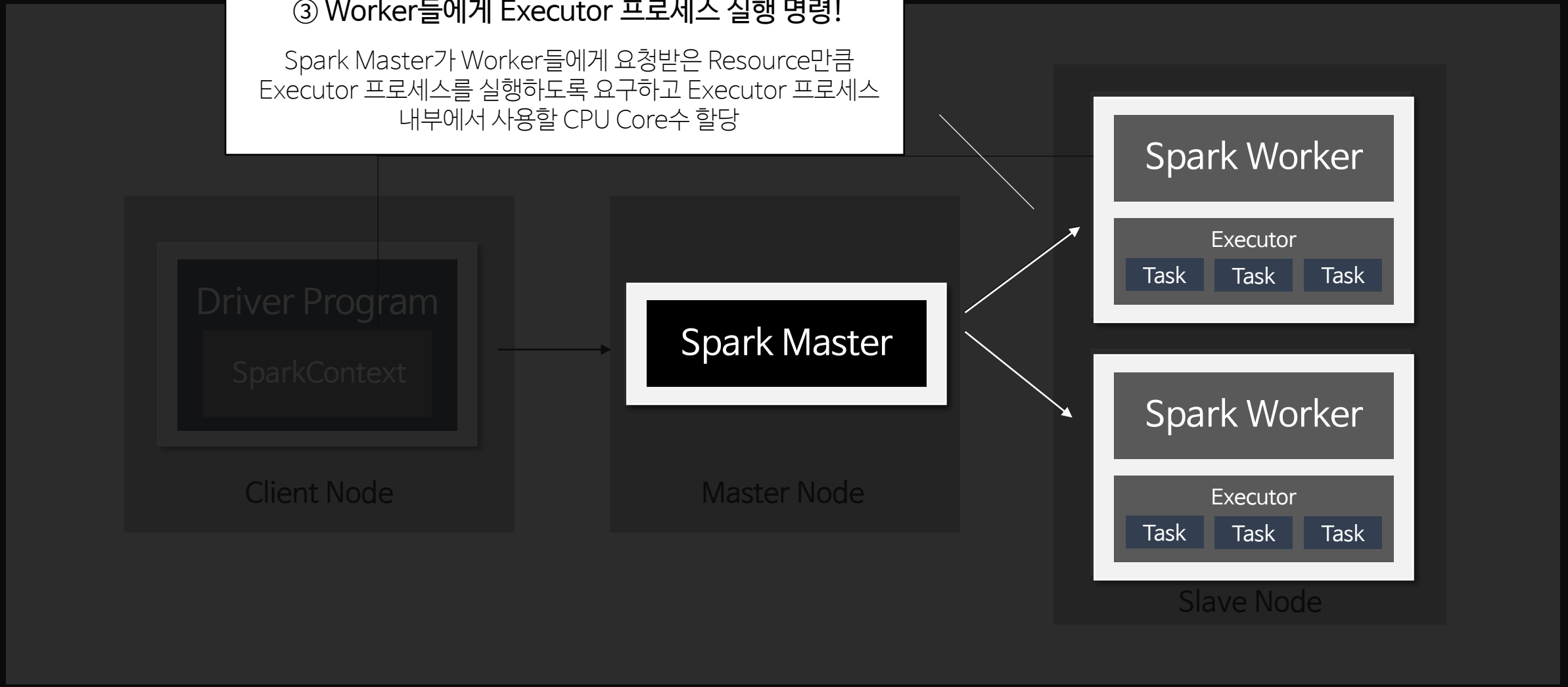


Spark?

- Spark 구성도 -

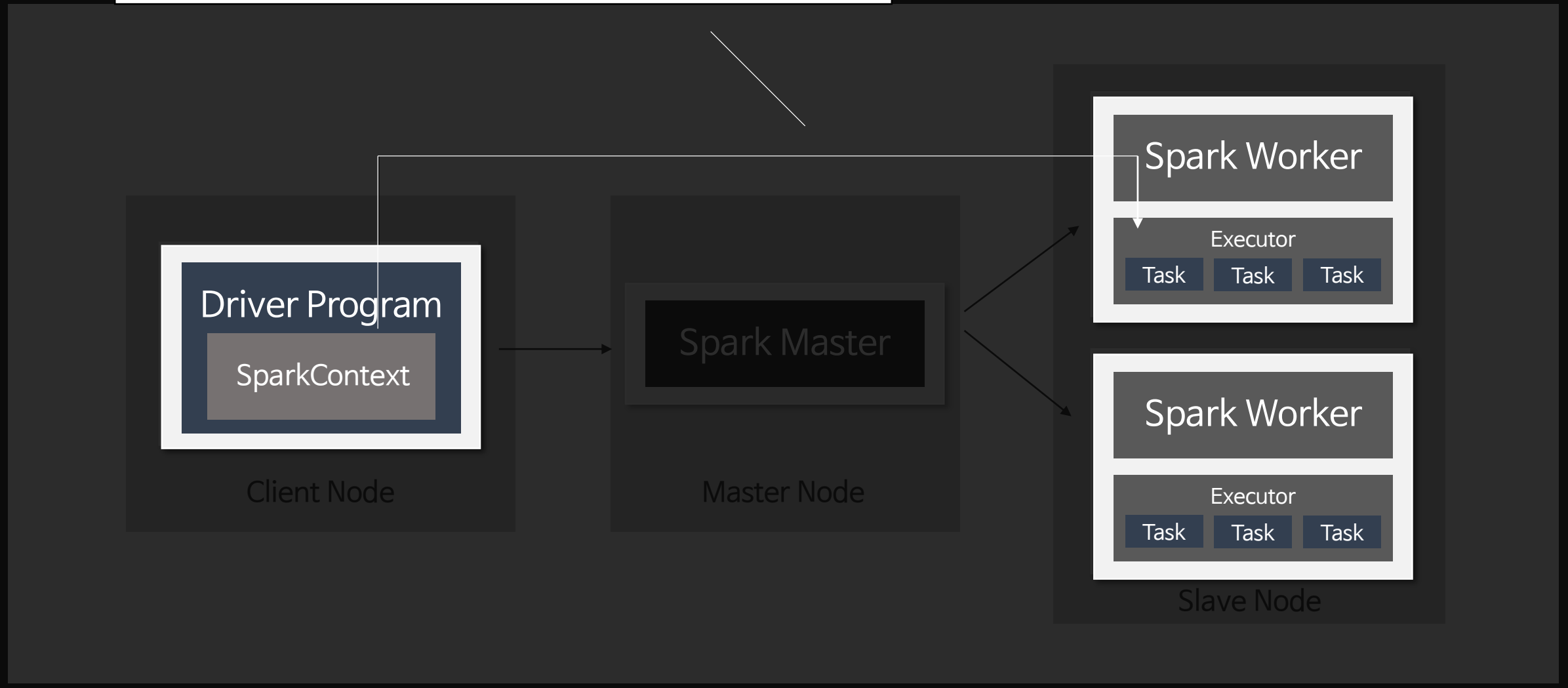
③ Worker들에게 Executor 프로세스 실행 명령!

Spark Master가 Worker들에게 요청받은 Resource만큼 Executor 프로세스를 실행하도록 요구하고 Executor 프로세스 내부에서 사용할 CPU Core수 할당



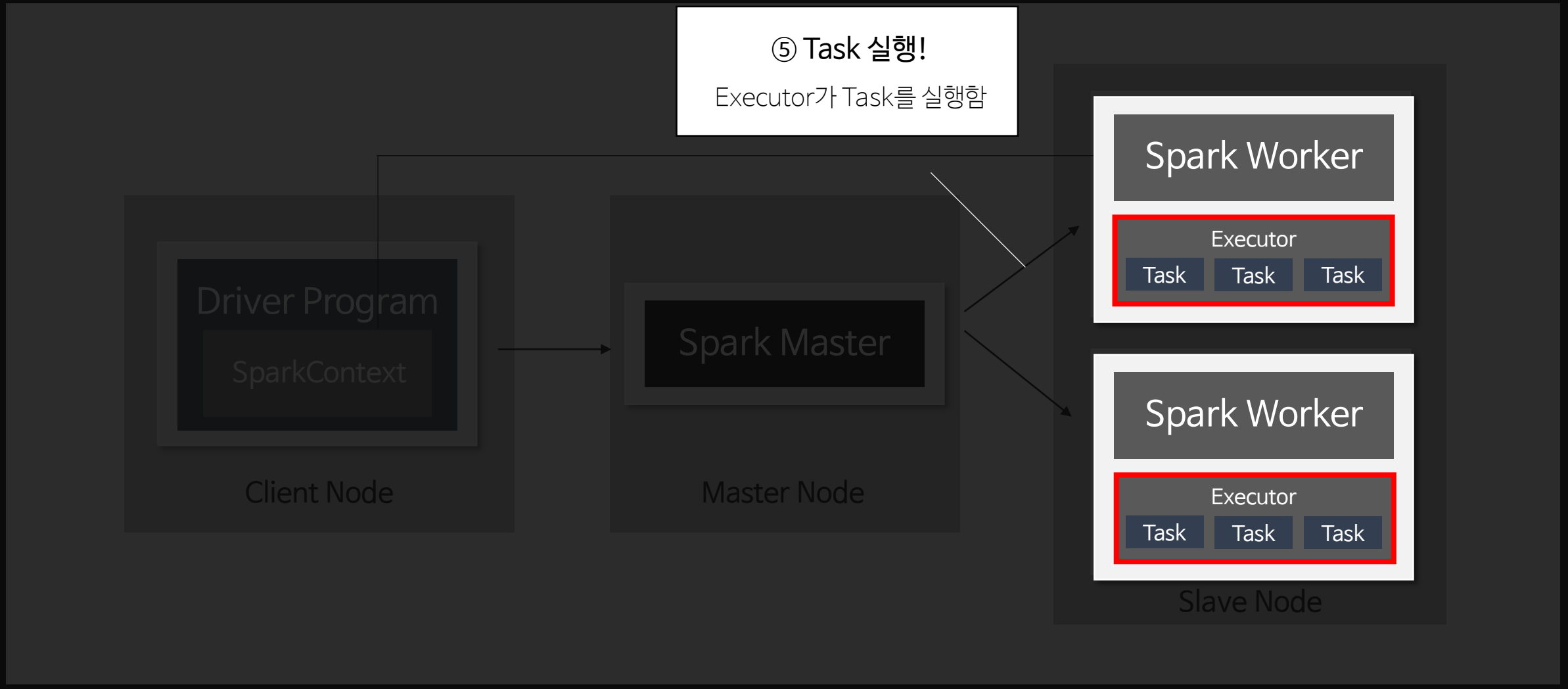
④ Task를 Executor에게 전송!

Driver Program이 Spark Application을 Task 단위로 나누어
Executor에게 전송함



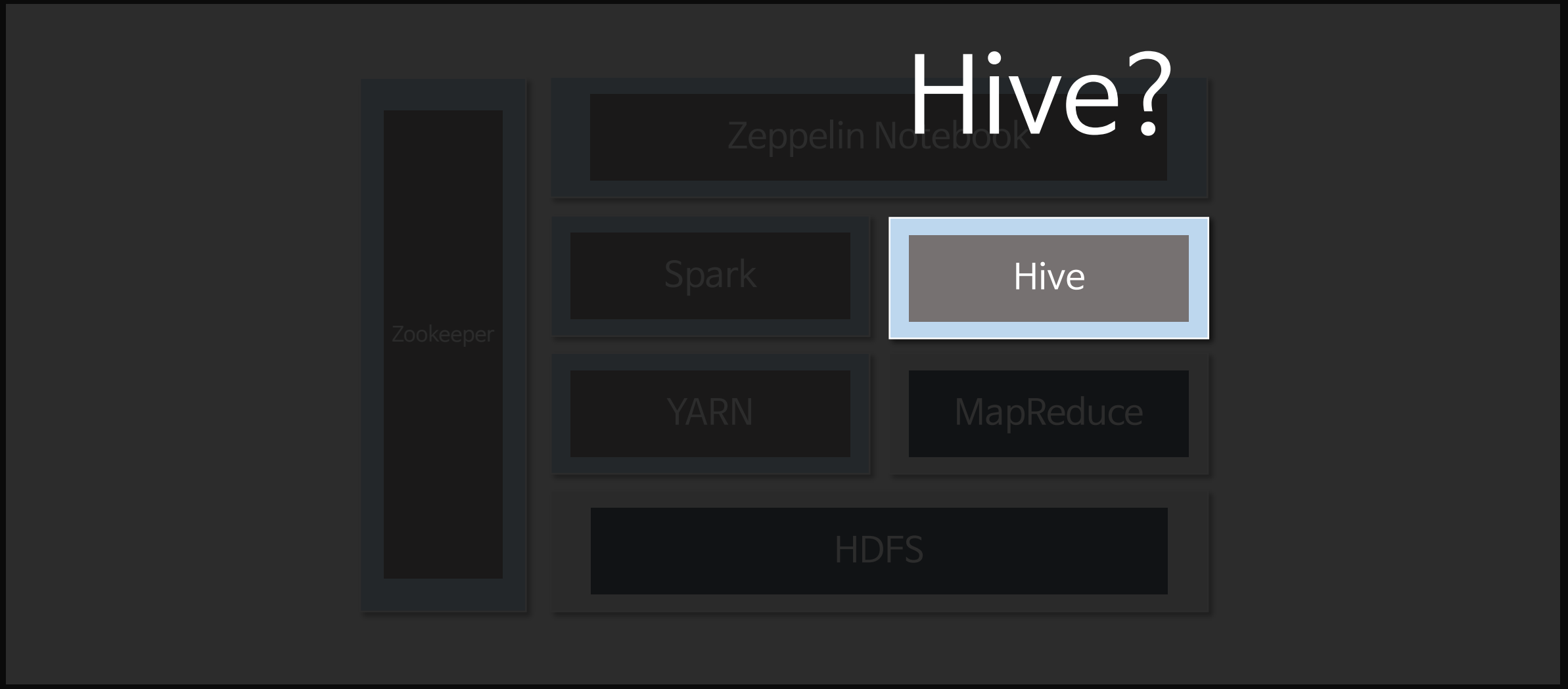
Spark?

- Spark 구성도 -



Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -

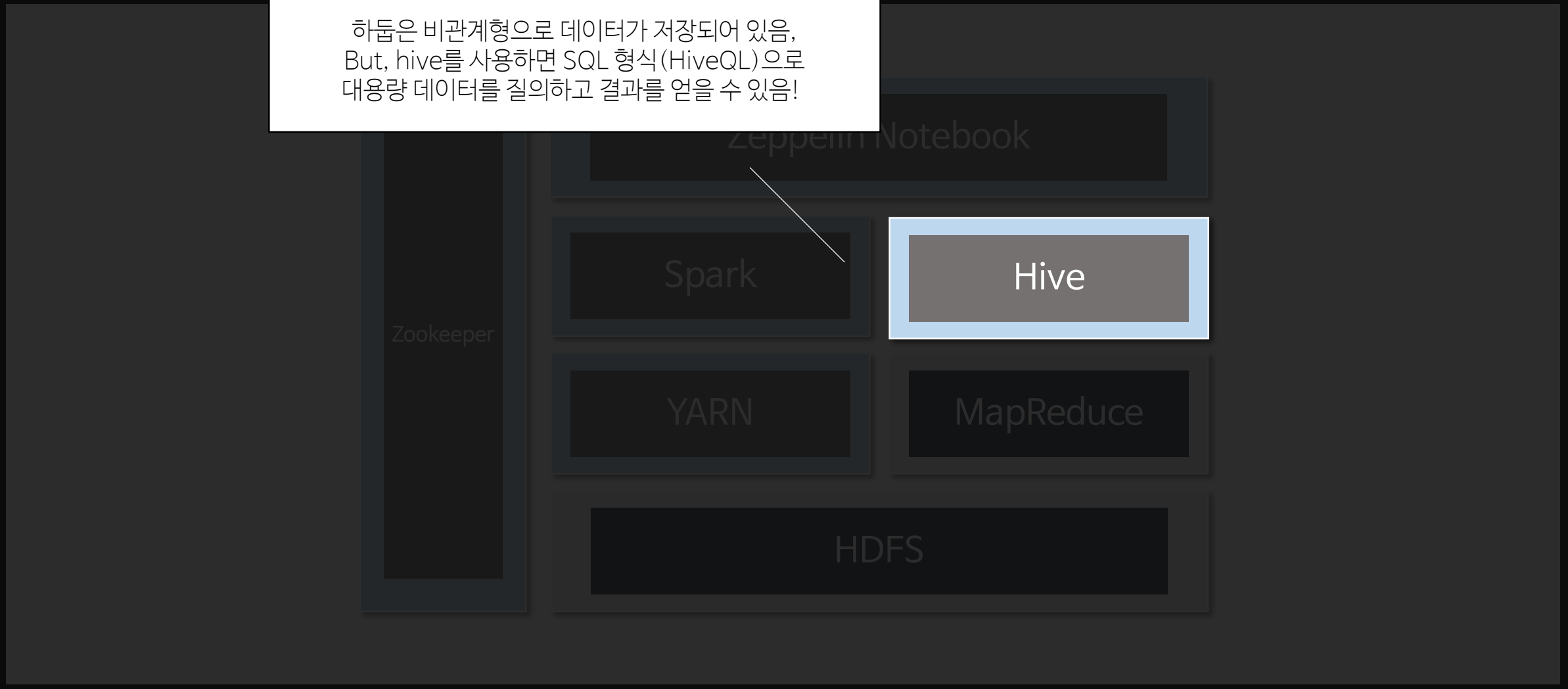


Hadoop ecosystem?

– 하둡 에코시스템이란 무엇일까? –

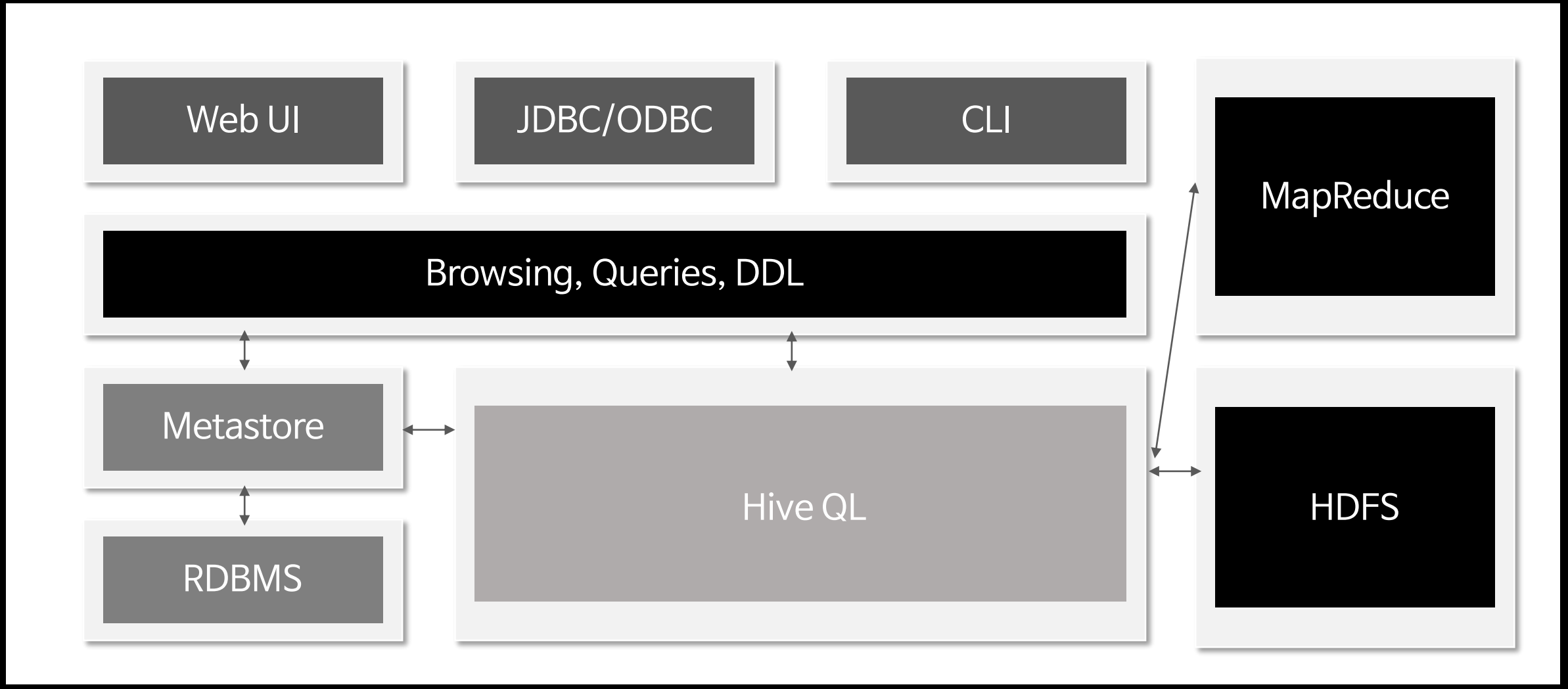
● 하둡 기반 SQL 쿼리 엔진!

하둡은 비관계형으로 데이터가 저장되어 있음,
But, hive를 사용하면 SQL 형식(HiveQL)으로
대용량 데이터를 질의하고 결과를 얻을 수 있음!



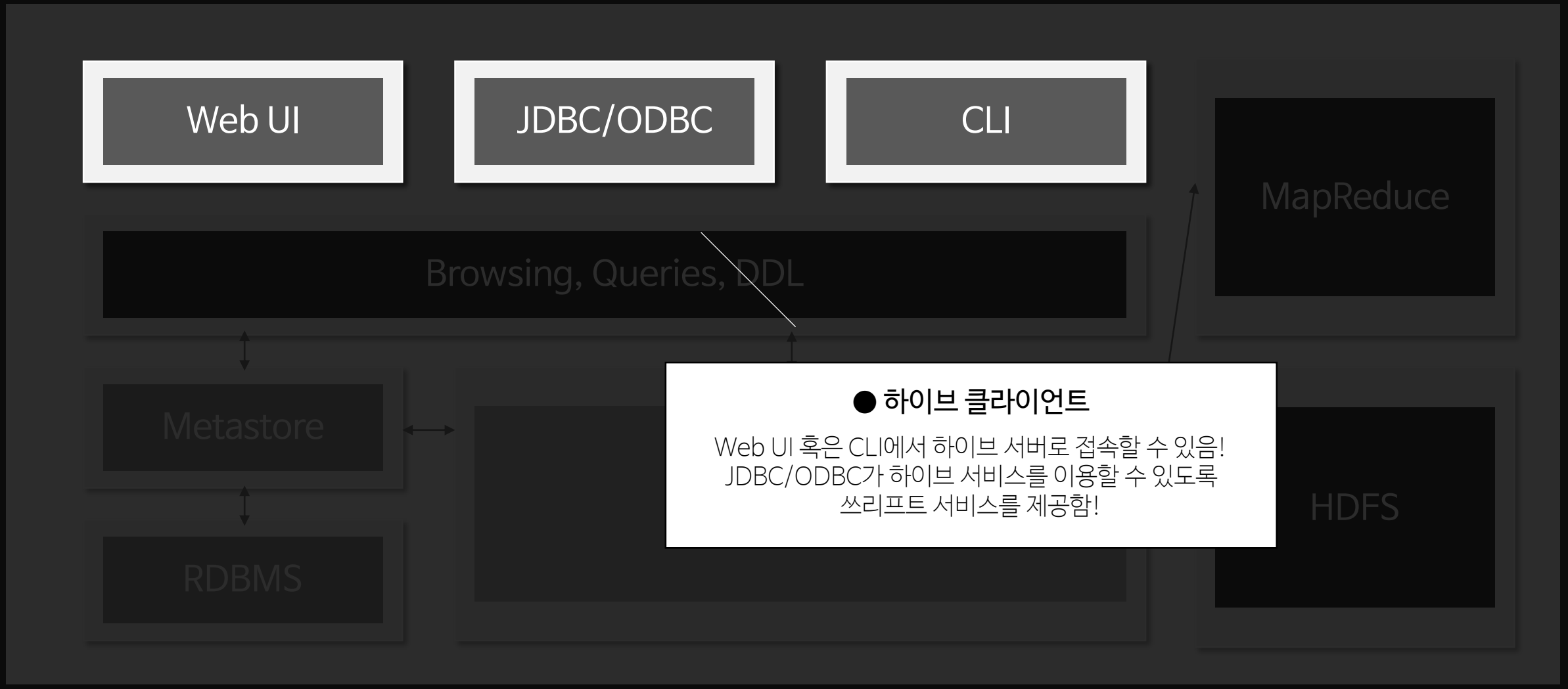
HIVE?

- HIVE의 구성도 -



HIVE?

- HIVE의 구성도 -

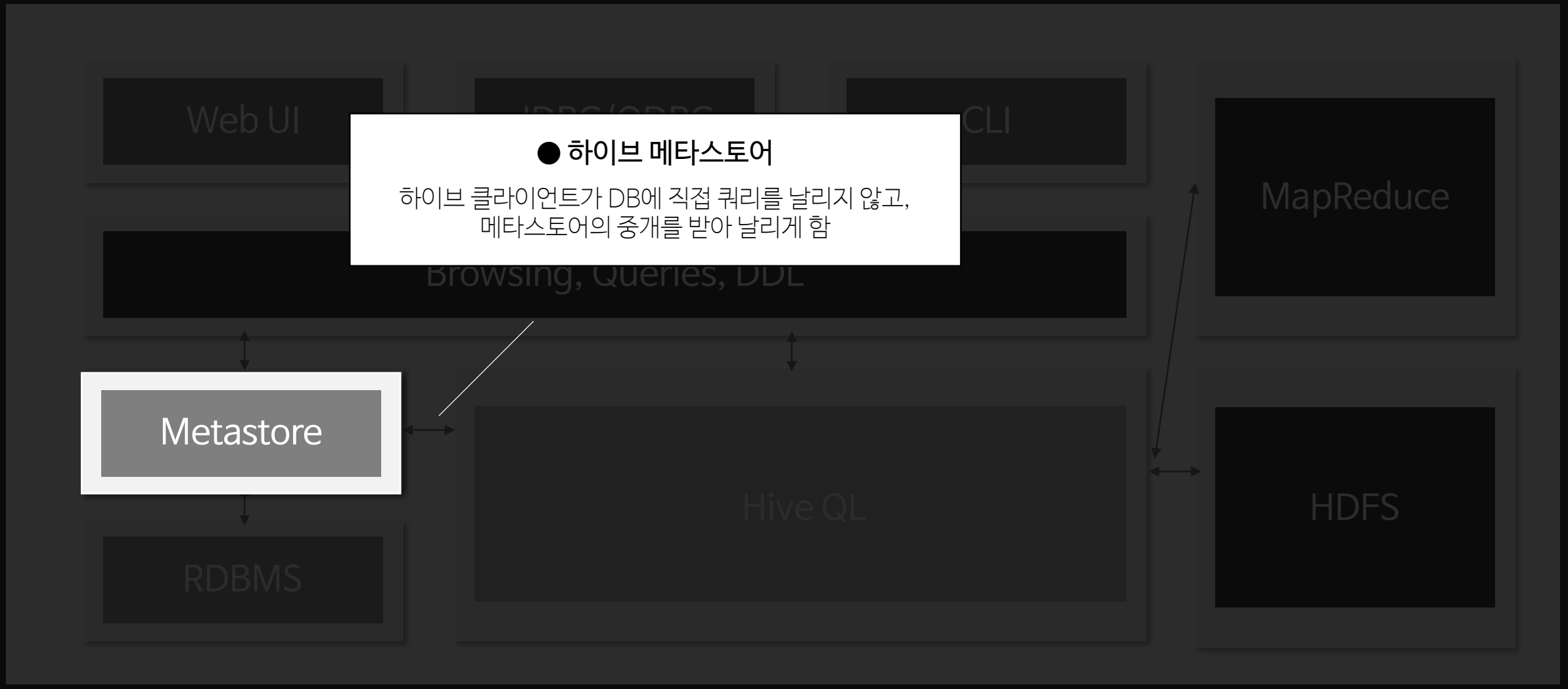


● 하이브 클라이언트

Web UI 혹은 CLI에서 하이브 서버로 접속할 수 있음!
JDBC/ODBC가 하이브 서비스를 이용할 수 있도록
쓰리프트 서비스를 제공함!

HIVE?

- HIVE의 구성도 -

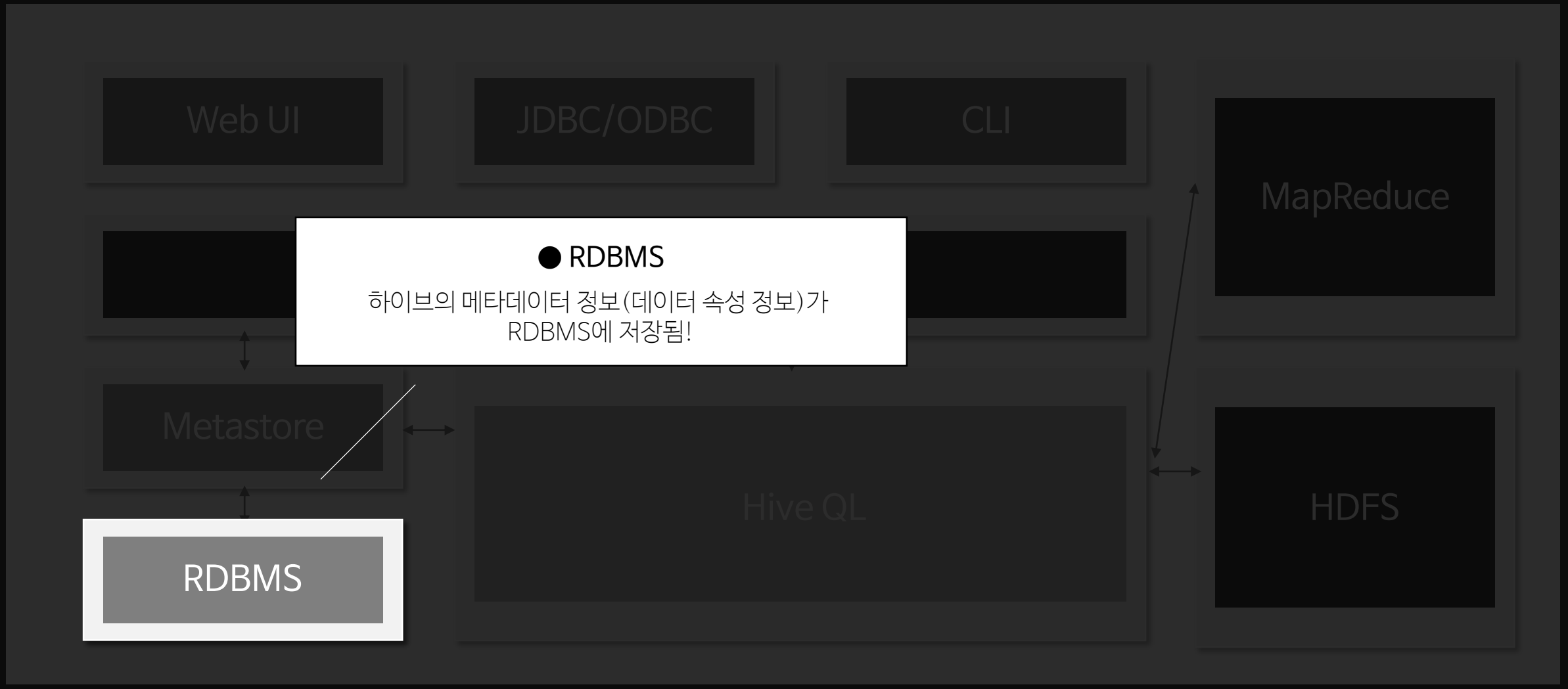


● 하이브 메타스토어

하이브 클라이언트가 DB에 직접 쿼리를 날리지 않고,
메타스토어의 중개를 받아 날리게 함

HIVE?

- HIVE의 구성도 -

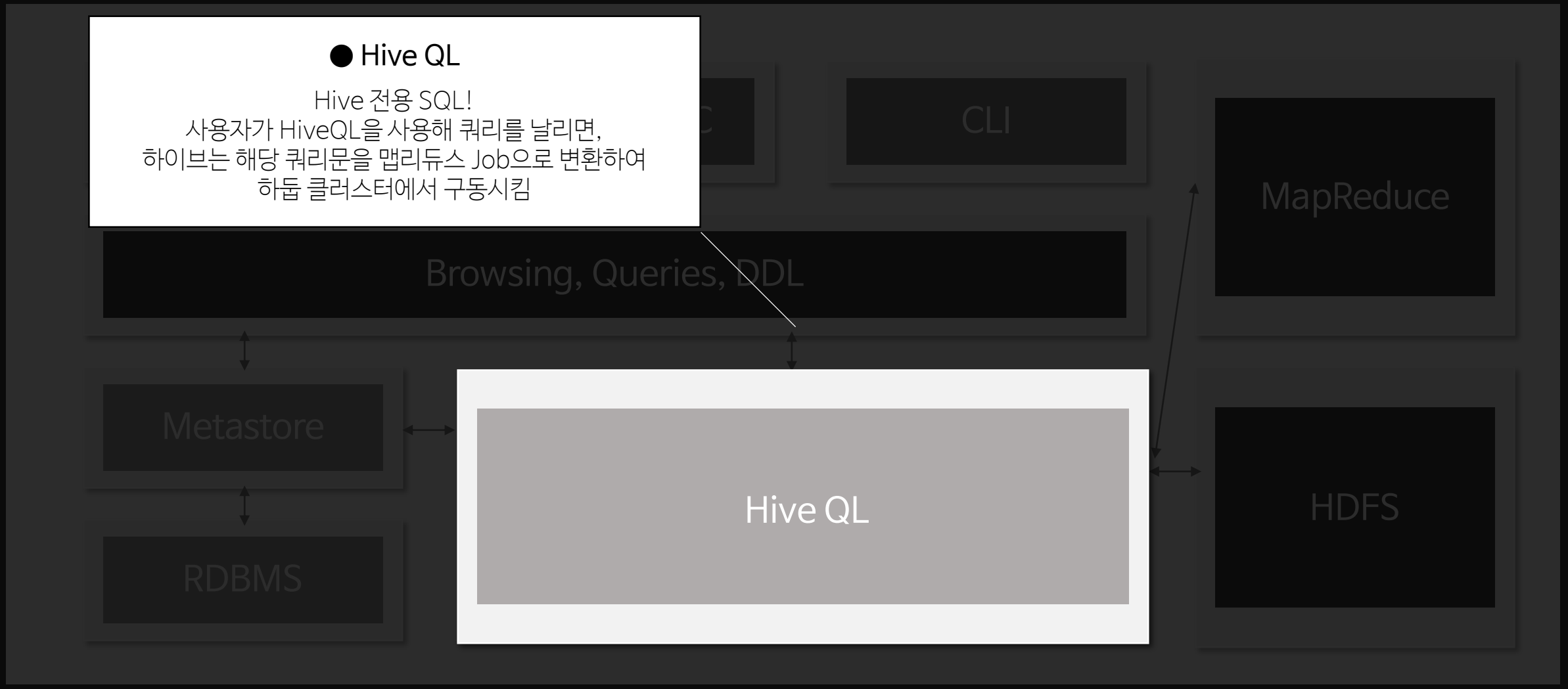


HIVE?

- HIVE의 구성도 -

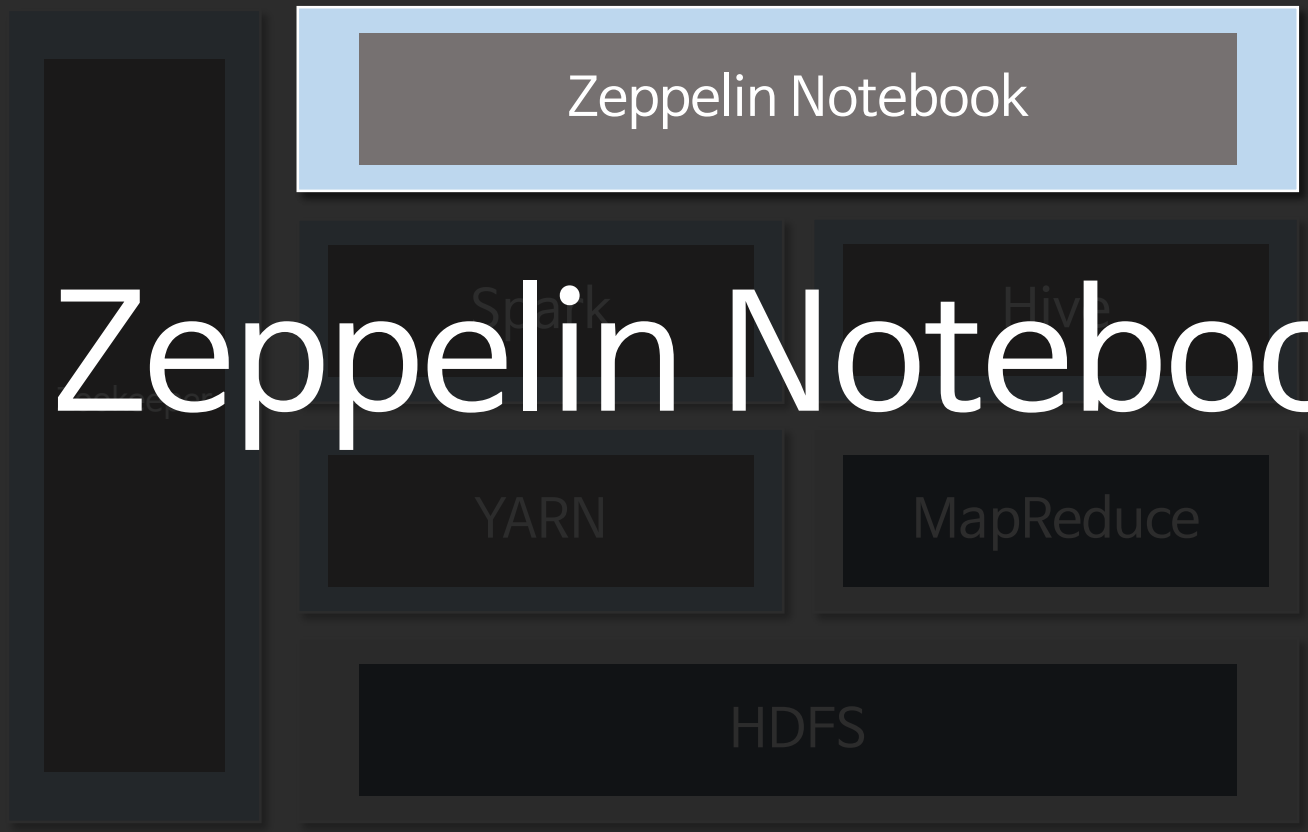
● Hive QL

Hive 전용 SQL!
사용자가 HiveQL을 사용해 쿼리를 날리면,
하이브는 해당 쿼리문을 맵리듀스 Job으로 변환하여
하둡 클러스터에서 구동시킴



Hadoop ecosystem?

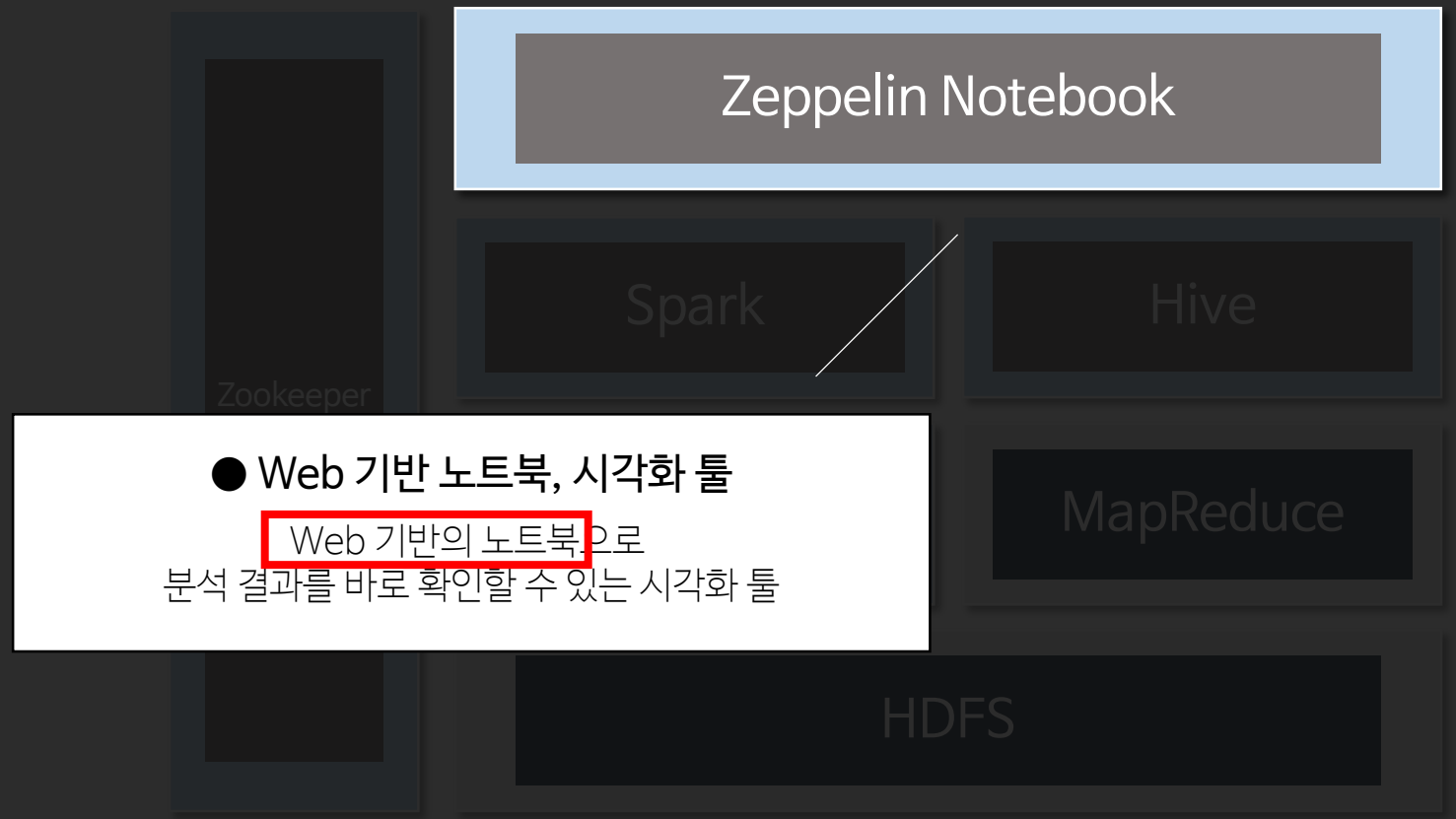
- 하둡 에코시스템이란 무엇일까? -



Zeppelin Notebook?

Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -

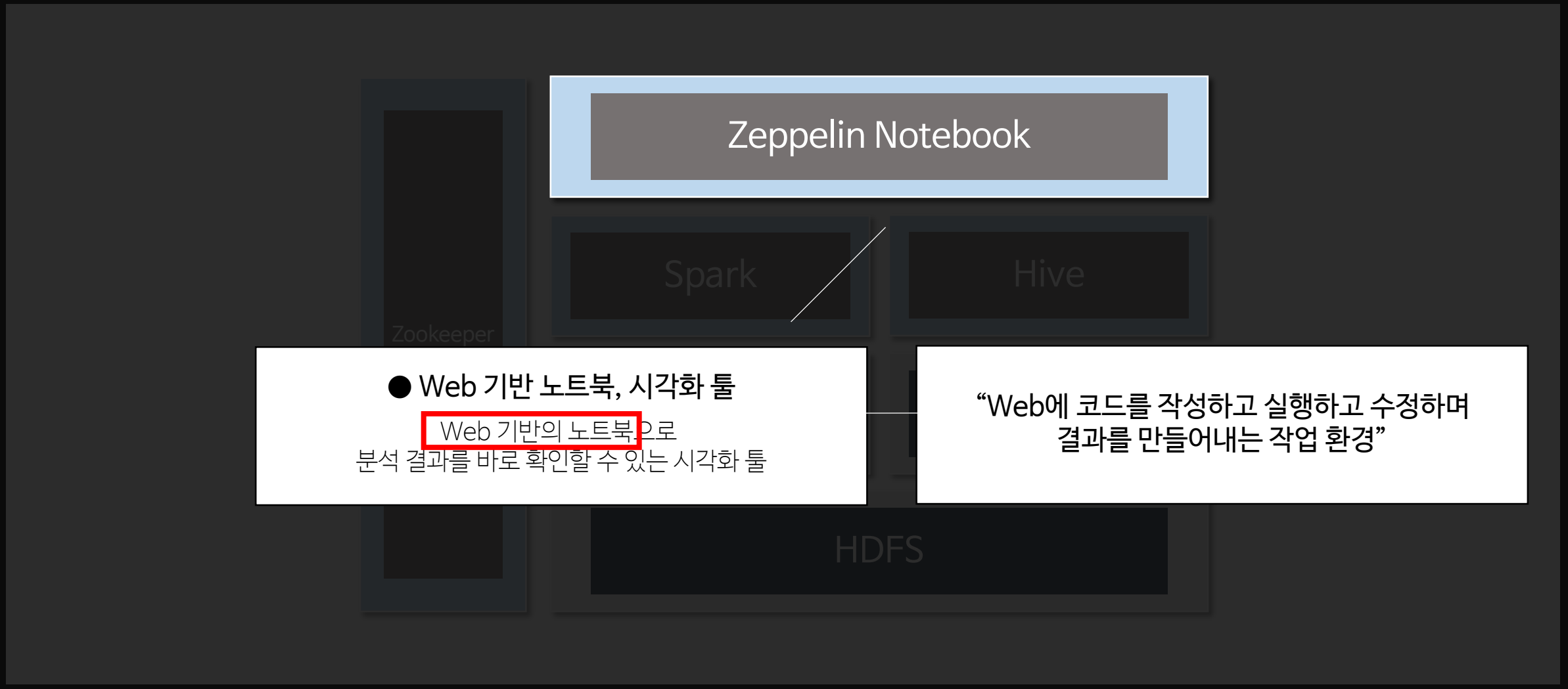


● Web 기반 노트북, 시각화 툴

Web 기반의 노트북으로
분석 결과를 바로 확인할 수 있는 시각화 툴

Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -



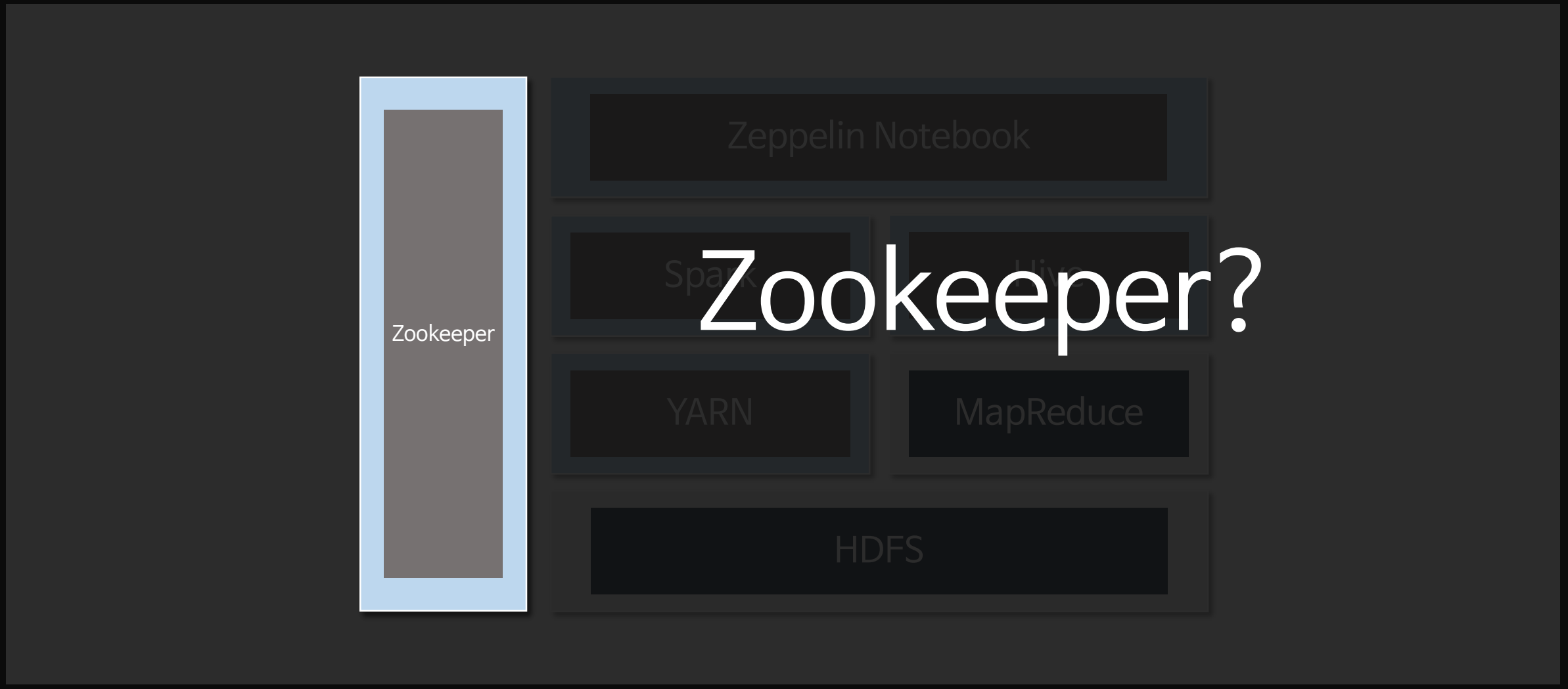
● Web 기반 노트북, 시각화 툴

Web 기반의 노트북으로
분석 결과를 바로 확인할 수 있는 시각화 툴

"Web에 코드를 작성하고 실행하고 수정하며
결과를 만들어내는 작업 환경"

Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -

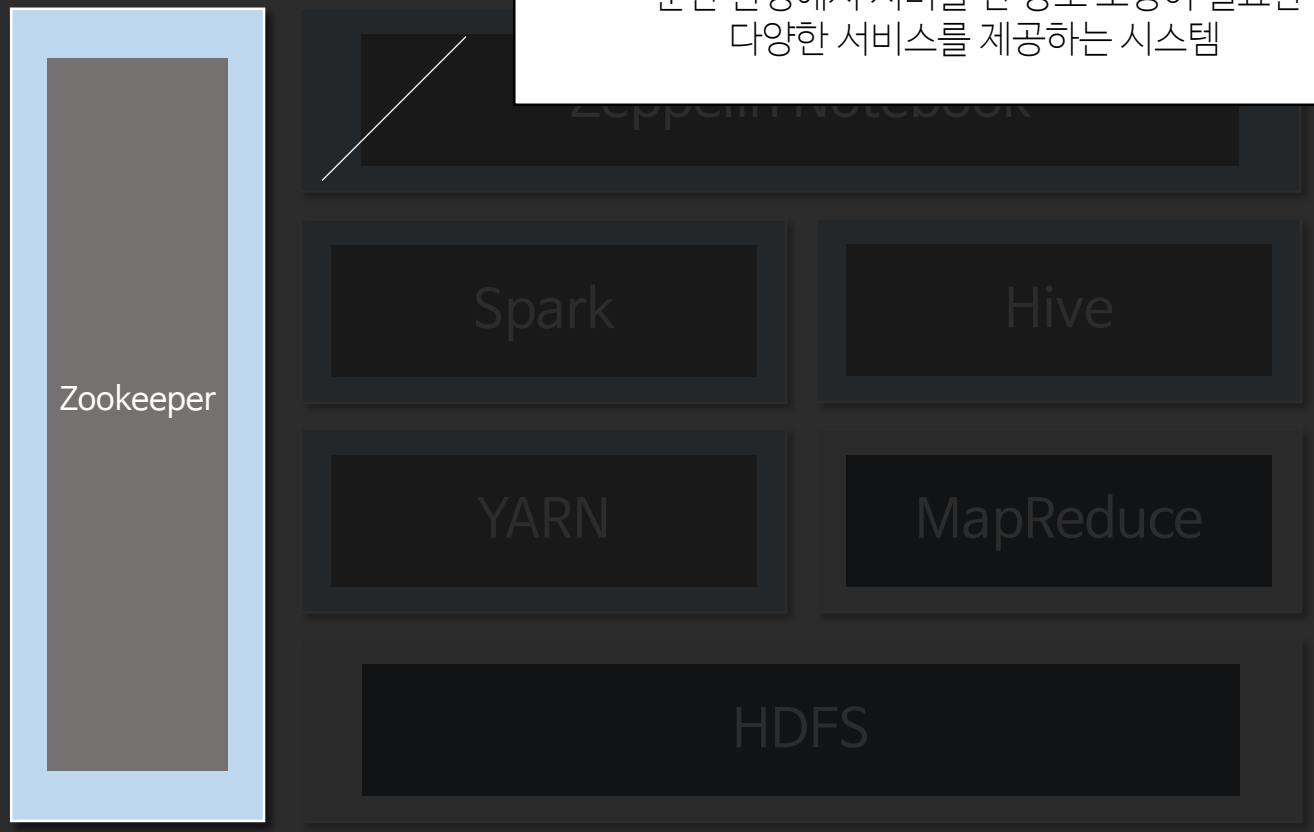


Hadoop ecosystem?

- 하둡 에코시스템이란 무엇일까? -

● 분산 시스템을 위한 코디네이터

분산 환경에서 서버들 간 상호 조정이 필요한
다양한 서비스를 제공하는 시스템



과제 내용

Condition A

NameNode가 설치된 PC와
다른 PC에 Hive를 설치할 것!

- 3개 이상의 DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것

  A. YARN / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  B. HDP를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  C. Data Block Size는 32MB로 설정
  D. Block Replication(복제)는 2로 설정
  E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공

  A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  JDBC를 제공할 것
  터가 존재하는 테이블을 3개 이상 만들 것

- Spark 서비스 제공
- Zeppelin notebook 서비스 제공

Hive Metastore는 Hive Server와
같은 PC에 존재해야 함!

과제 내용

Condition A

- DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것
  - / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  - 를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  - C. Data Block Size는 32MB로 설정
  - D. Block Replication(복제)는 2로 설정
  - E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공
  - A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  - B. Hive JDBC를 제공할 것
  - C. 데이터가 존재하는 테이블을 3개 이상 만들 것

- Spark 서비스 제공
- Zeppelin notebook 서비스 제공

JDBC를 반드시 설치하여
하이브 서비스를 이용할 수 있게 하자!

과제 내용

Condition A

- DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것
  - / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  - 를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  - C. Data Block Size는 32MB로 설정
  - D. Block Replication(복제)는 2로 설정
  - E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공
  - A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  - B. Hive JDBC를 제공할 것
  - C. 데이터가 존재하는 테이블을 3개 이상 만들 것

- Spark 서비스 제공
- Zeppelin notebook 서비스 제공

HiveQL로 질의를 날릴 테이블을
3개이상 만들어둘 것!

과제 내용

Condition A

- DataNode로 구성된 cluster를 구성하는 HDFS를 만들 것

  N / MR(MAP/REDUCE) / Zookeeper가 service로 존재할 것
  를 이용하는 경우 관리 DBMS는 MariaDB or Mysql
  C. Data Block Size는 32MB로 설정
  D. Block Replication(복제)는 2로 설정
  E. NameNode와 Hive service는 각각 다른 Host에서 구동할 것

- Hive 서비스 제공

  A. Hive metastore는 hive 서버(hive server master)와 동일한 호스트에 존재할 것
  B. Hive JDBC를 제공할 것
  C. 데이터가 존재하는 테이블을 3개 이상 만들 것

- Spark 서비스 제공
- Zeppelin notebook 서비스 제공

Condition B

과제 내용

Condition B

- 각 Host는 아래의 사양을 만족해야 한다.
  - A. 각 Host의 Disk Size는 48 ~ 64GB로 설정한다.
  - B. 각 Host의 CPU Core는 2~4개로 설정한다.
  - C. 각 Host의 RAM Size는 4.8GB 이상으로 설정한다.

- Root는 패스워드가 없어야 한다.
  - A. HDP 설치 시 ssh private key 입력을 통한 인증(키교환)으로 Host를 등록해야 한다.

- 클러스터는 외부에서 접속 가능하도록 가상 브릿지 네트워크를 이용한 통신 구성을 해야한다.
  - A. 네트워크 대역은 상관 없음

- 방화벽 서비스는 disable
- 서비스 연결 시 host명 return 문제를 해결해야 한다.

SSH 인증방식?

- SSH 인증방식이란 무엇일까? -

"원격접속을 위한 프로토콜"

SSH 인증방식?

- SSH 인증방식이란 무엇일까? -

"기존의 ID/PW 접속 방식이 아닌 키 인증을 통한 접속 "

Private key

Public key

Public key

ssh-keygen 을 통해 키를 생성하면
Private key(개인키), Public key(공개키)가
생성됨!

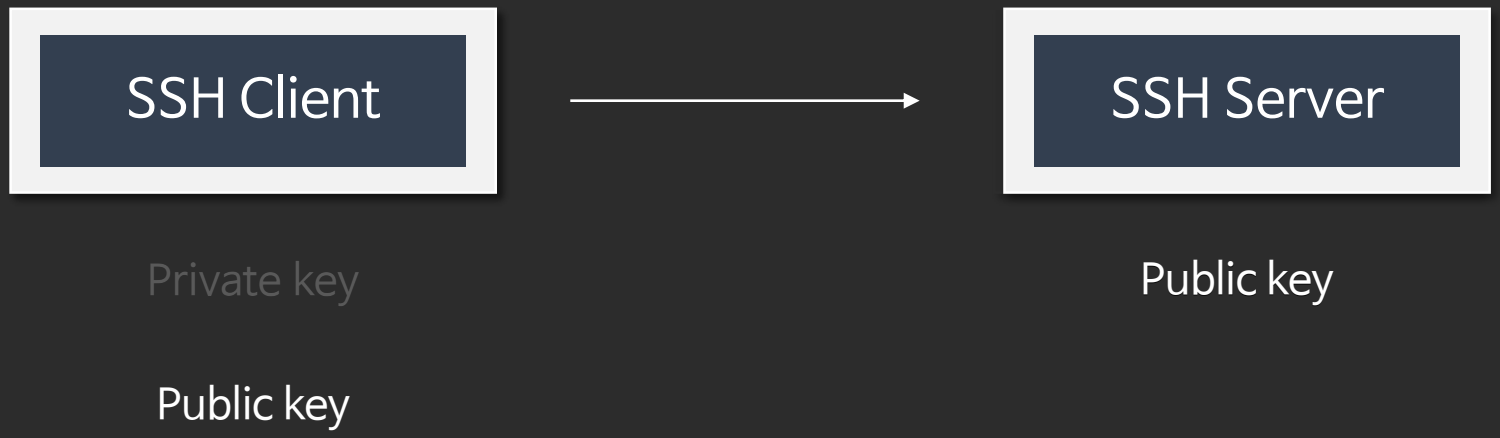
“기존의 ID/PW 접속 방식이 아닌 키 인증을 통한 접속 "

SSH 인증방식?

- SSH 인증방식이란 무엇일까? -

Public key(공개키)를 원격 접속할 컴퓨터에
복사하면 공개키 인증을 통해 패스워드 없이
원격 접속 가능!

"기존의 ID 한 접속 "



과제 내용

Condition B

- 각 Host는 아래의 사양을 만족해야 한다.
  - A. 각 Host의 Disk Size는 48 ~ 64GB로 설정한다.
  - B. 각 Host의 CPU Core는 2~4개로 설정한다.
  - C. 각 Host의 RAM Size는 4.8GB 이상으로 설정한다.

- Root는 패스워드가 없어야 한다.
  - A. HDP 설치 시 ssh private key 입력을 통한 인증(키교환)으로 Host를 등록해야 한다.

- 클러스터는 외부에서 접속 가능하도록 가상 브릿지 네트워크를 이용한 통신 구성을 해야한다.
  - A. 네트워크 대역은 상관 없음

- 방화벽 서비스는 disable
- 서비스 연결 시 host명 return 문제를 해결해야 한다.

SSH 키 인증방식으로 패스워드 없이 접속할 수 있도록 클러스터를 구성해라!

과제 내용

Condition B

- 각 Host는 아래의 사양을 만족해야 한다.
  - A. 각 Host의 Disk Size는 48 ~ 64GB로 설정한다.
  - B. 각 Host의 CPU Core는 2~4개로 설정한다.
  - C. 각 Host의 RAM Size는 4.8GB 이상으로 설정한다.

- Root는 패스워드가 없어야 한다.
  - A. HDP 설치 시 ssh private key 입력을 통한 인증(키교환)으로 Host를 등록해야 한다.

- 클러스터는 외부에서 접속 가능하도록 가상 브릿지 네트워크를 이용한 통신 구성을 해야한다.
  - A. 네트워크 대역은 상관 없음

- 방화벽 서비스는 disable
- 서비스 연결 시 host명 return 문제를 해결해야 한다.

가상 브릿지 네트워크?

- 가상 브릿지 네트워크란 무엇일까? -

공유기로부터 IP를 할당 받아, 각 가상머신이 호스트 PC와 같은 대역의 IP를 갖게 되는 방식

가상 브릿지 네트워크?

- 가상 브릿지 네트워크란 무엇일까? -

공유기를 통해 외부 네트워크와의 통신이 가능해짐!

과제 내용

- 각 Host는 아래의 사양을 만족해야 한다.
  - A. 각 Host의 Disk Size는 48 ~ 64GB로 설정한다.
  - B. 각 Host의 CPU Core는 2~4개로 설정한다.
  - C. 각 Host의 RAM Size는 4.8GB 이상으로 설정한다.
- Root는 패스워드가 없어야 한다.
  - A. HDP 설치 시 ssh private key 입력을 통한 인증(키교환)으로 Host를 등록해야 한다.
- 는 외부에서 접속 가능하도록 가상 브릿지 네트워크를 이용한 통신 구성을 해야한다.
  - A. 네트워크 대역은 상관 없음
- 방화벽 서비스는 disable
- 서비스 연결 시 host명 return 문제를 해결해야 한다.

방화벽이 설정되있을 경우 외부에서 접속이 안되므로 disable로 설정해야함!